



















CURRENT YR/VOL

Marygrove College Library  
8425 West McNichols Road  
Detroit, MI 48221

Volume 100  
Number 1

February 2008

Published quarterly  
by the  
American Psychological  
Association

ISSN 0022-0663

# Journal of Educational Psychology

Karen R. Harris, *Editor*

Eric M. Anderman, *Associate Editor*

Donna M. Kulikowich, *Associate Editor*

Gloria Miller, *Associate Editor*

Frank Pajares, *Associate Editor*

Jeffrey J. Walczyk, *Associate Editor*



## Editor

Karen R. Harris, *Vanderbilt University*

## Associate Editors

Eric M. Anderman, *Ohio State University*  
Jonna M. Kulikowich, *Pennsylvania State University*  
Gloria Miller, *University of Denver*  
Frank Pajares, *Emory University*  
Jeffrey J. Walczyk, *Louisiana Technical University*

## Chief Editorial Assistant

Brenna Hansen, *Vanderbilt University*

## Editorial Assistants

Karrie Godwin, *University of Denver*  
Diana Griffith-Ross, *Louisiana Technical University*  
Jason Chen, *Emory University*  
Nicholas D. Warcholak, *Pennsylvania State University, University Park Campus*

## Advisory Editors

Patricia Alexander, *University of Maryland, College Park*  
Ellen R. Altermatt, *Hanover College*  
Lynley H. Anderman, *Ohio State University*  
Robert Atkinson, *Arizona State University*  
Carole Beal, *Information Sciences Institute at the University of Southern California*  
Hefer Bembenutty, *Queens College*  
David A. Bergin, *University of Missouri—Columbia*  
Benita A. Blachman, *Syracuse University*  
Mimi Bong, *Ewha Womans University, Seoul, Korea*  
Jere Brophy, *Michigan State University*  
Scott W. Brown, *University of Connecticut*  
Adriana G. Bus, *Leiden University, Leiden, the Netherlands*  
Robert Calfee, *University of California, Riverside*  
Joanne F. Carlisle, *University of Michigan*  
Martha Carr, *University of Georgia*  
Jerrold C. Cassidy, *Ball State University*  
Clark Chinn, *Rutgers University*  
Namok Choi, *University of Louisville*  
Donald L. Compton, *Vanderbilt University*  
Alice J. Corkill, *University of Nevada, Las Vegas*  
H. Michael Crowson, *University of Oklahoma*  
Anne E. Cunningham, *University of California, Berkeley*  
Teresa K. DeBacker, *University of Oklahoma*  
Amanda M. Durik, *Northern Illinois University*  
Pamela Beard El-Dinary, *Educational Consultant*  
Dorothy L. Espelage, *University of Illinois at Urbana—Champaign*  
Jill Fitzgerald, *University of North Carolina at Chapel Hill*  
Douglas Fuchs, *Vanderbilt University*  
Lynn S. Fuchs, *Vanderbilt University*  
David C. Geary, *University of Missouri*  
Alexandra Gottardo, *Wilfrid Laurier University, Waterloo, Ontario, Canada*  
Steve Graham, *Vanderbilt University*  
Barbara A. Greene, *University of Oklahoma*  
Charles R. Greenwood, *University of Kansas*  
John Guthrie, *University of Maryland, College Park*  
Douglas J. Hacker, *University of Utah*  
Vernon C. Hall, *Syracuse University*  
Jenefer Husman, *Arizona State University*  
Michael L. Kamil, *Stanford University*  
Avi Kaplan, *Ben Gurion University of the Negev, Beer Sheva, Israel*  
Robert M. Klassen, *University of Alberta, Edmonton, Alberta, Canada*  
Beth Kurtz-Costes, *University of North Carolina at Chapel Hill*  
Dan Lapsley, *University of Notre Dame*  
Steve Lehman, *Utah State University*  
Willy Lens, *University of Leuven, Leuven, Belgium*  
Joel R. Levin, *University of Arizona*  
Elizabeth A. Linnenbrink, *Duke University*  
Mary Lundeborg, *Michigan State University*  
Charles MacArthur, *University of Delaware*  
Linda H. Mason, *Pennsylvania State University, University Park Campus*  
Richard E. Mayer, *University of California, Santa Barbara*  
Catherine McBride-Chang, *Chinese University of Hong Kong, Shatin, Hong Kong, China*  
Valentina McInerney, *University of Western Sydney*  
Debra K. Meyer, *Elmhurst College*  
Michael Middleton, *University of New Hampshire*  
Lisa M. Soederberg Miller, *University of California, Davis*  
Raymond B. Miller, *University of Oklahoma*  
Jens Möller, *University of Kiel, Kiel, Germany*  
Tamera B. Murdock, *University of Missouri—Kansas City*  
Karen P. Murphy, *Pennsylvania State University, University Park Campus*  
Darcia Narvaez, *University of Notre Dame*  
Markku Niemivirta, *University of Helsinki, Helsinki, Finland*  
Jane Oakhill, *University of Sussex, Falmer, Brighton, United Kingdom*  
Rollanda E. O'Connor, *University of California, Riverside*  
Richard Olson, *University of Colorado*  
Helen Patrick, *Purdue University*  
Nancy Perry, *University of British Columbia, Vancouver, British Columbia, Canada*  
Gary Phye, *Iowa State University*  
Jan L. Plass, *New York University*  
Robert Reid, *University of Nebraska—Lincoln*  
Robert Renaud, *University of Manitoba, Winnipeg, Manitoba, Canada*  
Alison M. Ryan, *University of Illinois at Urbana—Champaign*  
Hollis S. Scarborough, *Haskins Laboratories, New Haven, Connecticut*  
Christopher Schatschneider, *Florida State University*  
Wolfgang Schneider, *Universität Würzburg, Würzburg, Germany*  
Marlene Schommer-Aikins, *Wichita State University*  
Gregory Schraw, *University of Nevada, Las Vegas*

Einar M. Skaalvik, *Norwegian University of Science and Technology, Trondheim, Norway*  
Susan Sonnenschein, *University of Maryland, Baltimore County*  
Laura M. Stapleton, *University of Maryland, Baltimore County*  
Joseph Stevens, *University of Oregon*  
H. Lee Swanson, *University of California, Riverside*  
John Sweller, *University of New South Wales, Sydney, New South Wales, Australia*  
Sonya Symons, *Acadia University, Wolfville, Nova Scotia, Canada*  
Keith Thiede, *University of Illinois at Chicago*  
Theresa A. Thorkildsen, *University of Illinois at Chicago*  
Tim Urdan, *Santa Clara University*  
Ellen Usher, *University of Kentucky*  
Giovanni Valiante, *Rollins College*  
Sharon Vaughn, *University of Texas at Austin*  
Regina Vollmeyer, *University of Frankfurt, Frankfurt, Germany*  
Charles A. Weaver III, *Baylor University*  
Kathryn R. Wentzel, *University of Maryland, College Park*  
Allan Wigfield, *University of Maryland, College Park*  
Joanna P. Williams, *Teachers College, Columbia University*  
Christopher A. Wolters, *University of Houston*  
Moshe Zeidner, *University of Haifa, Haifa, Israel*  
Barry J. Zimmerman, *Graduate Center, City University of New York*

The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

**Change of Address:** Send change of address notice and a recent mailing label to the attention of Subscriptions Department, APA, 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee periodicals forwarding postage.

**Single Issues, Back Issues, and Back Volumes:** For information regarding single issues, back issues, or back volumes, write to Order Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

**Microform Editions:** For information regarding microform editions, write to University Microfilms, Ann Arbor, MI 48106.

**Manuscripts:** Submit manuscripts electronically through the Manuscript Submissions Portal found at [www.apa.org/journals/edu](http://www.apa.org/journals/edu), according to the Instructions to Authors found elsewhere in this issue (see table of contents). Correspondence regarding manuscripts should be sent to the Editor, Karen R. Harris, Vanderbilt University, *Journal of Educational Psychology*, Box 507 Peabody College, Nashville, Tennessee 37203-5721. The opinions and statements published are the responsibility of the authors, and such opinions and statements do not necessarily represent the policies of APA or the views of the Editor.

**Copyright and Permission:** Those who wish to reuse APA-copyrighted material in a non-APA publication must secure from APA and the author of the material written permission to reproduce a journal article in full or journal text of more than 500 words. APA normally grants permission contingent on permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of a fee of \$25 per page. Permission from APA and fees are waived for those who wish to reproduce a single table or figure from a journal for use in a print product, provided the author's permission is obtained and full credit is given to APA as copyright holder and to the author through a complete citation. (Requesters requiring written permission for commercial use of a single table or figure will be assessed a \$25 service fee.) Permission and fees are waived for authors who wish to reproduce their own material for personal use; fees only are waived for authors who wish to use more than a single table or figure of their own material commercially (but for use in edited books, fees are waived for the author only if serving as the book editor). Permission and fees are waived for the photocopying of isolated journal articles for nonprofit classroom or library reserve use by instructors and educational institutions. A permission fee may be charged to the requester if students are charged for the material, multiple articles are copied, or large-scale copying is involved (e.g., for course packs). Access services may use unedited abstracts without the permission of APA or the author. Libraries are permitted to photocopy beyond the limits of the U.S. copyright law: (1) post-1977 articles, provided the per-copy fee in the code for this journal (0022-0663/08/\$12.00) is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923; (2) pre-1978 articles, provided that the per-copy fee stated in the Publishers' Fee List is paid through the Copyright Clearance Center. Address requests for reprint permission to Permissions Office, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

**Electronic Access:** APA members who subscribe to this journal have automatic access to a 3-year file of the journal in the PsycARTICLES® full-text database. See <http://member-s.apa.org/access>.

**Reprints:** Authors may order reprints of their articles from the printer when they receive proofs.

**APA Journal Staff:** Susan J. A. Harris, *Senior Director, Journals Program*; Skip Maier, *Director, Journal Services*; Paige W. Jackson, *Director, Editorial Services*; Clark Munsell, *Account Manager*; Jodi Ashcraft, *Advertising Sales Manager*.

The **Journal of Educational Psychology** (ISSN 0022-0663) is published quarterly (February, May, August, November) in one volume per year by the American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Subscriptions are available on a calendar year basis only (January through December). The 2008 rates follow: *Nonmember Individual*: \$161 Domestic, \$185 Foreign, \$195 Air Mail. *Institutional*: \$450 Domestic, \$491 Foreign, \$504 Air Mail. *APA Member*: \$73. Write to Subscriptions Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Printed in the U.S.A. Periodicals postage paid at Washington, DC, and at additional mailing offices. POSTMASTER: Send address changes to the *Journal of Educational Psychology*, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.



# Educational Psychology

February 2008

Volume 100

Number 1

[www.apa.org/journals/edu](http://www.apa.org/journals/edu)

## Articles

Copyright © 2008  
by the  
American  
Psychological  
Association

- 1 Teacher–Student Support, Effortful Engagement, and Achievement: A 3-Year Longitudinal Study  
*Jan N. Hughes, Wen Luo, Oi-Man Kwok, and Linda K. Loyd*
- 15 Interplay Between Personal Goals and Classroom Goal Structures in Predicting Student Outcomes: A Multilevel Analysis of Person–Context Interactions  
*Shun Lau and Youyan Nie*
- 30 Problem Solving and Computational Skill: Are They Shared or Distinct Aspects of Mathematical Cognition?  
*Lynn S. Fuchs, Douglas Fuchs, Karla Stuebing, Jack M. Fletcher, Carol L. Hamlett, and Warren Lambert*
- 48 Mathematical Competencies in Children With Different Types of Learning Difficulties  
*Ulf Andersson*
- 67 Prediction of Children’s Academic Competence From Their Effortful Control, Relationships, and Classroom Participation  
*Carlos Valiente, Kathryn Lemery-Chalfant, Jodi Swanson, and Mark Reiser*
- 78 A Multilevel Perspective on Gender in Classroom Motivation and Climate: Potential Benefits of Male Teachers for Boys?  
*Herbert W. Marsh, Andrew J. Martin, and Jacqueline H. S. Cheng*
- 96 A Multilevel Study of Predictors of Student Perceptions of School Climate: The Effect of Classroom-Level Factors  
*Christine W. Koth, Catherine P. Bradshaw, and Philip J. Leaf*
- 105 The Role of Achievement Goals in the Development of Interest: Reciprocal Relations Between Achievement Goals, Interest, and Performance  
*Judith M. Harackiewicz, Amanda M. Durik, Kenneth E. Barron, Lisa Linnenbrink-Garcia, and John M. Tauer*
- 123 A Response to Recent Reanalyses of the National Reading Panel Report: Effects of Systematic Phonics Instruction Are Practically Significant  
*Karla K. Stuebing, Amy E. Barth, Paul T. Cirino, David J. Francis, and Jack M. Fletcher*
- 135 Text Comprehension in Chinese Children: Relative Contribution of Verbal Working Memory, Pseudoword Reading, Rapid Automatized Naming, and Onset-Rime Phonological Segmentation  
*Che Kan Leong, Shek Kam Tse, Ka Yee Loh, and Kit Tai Hau*
- 150 Development of Word Reading Fluency and Spelling in a Consistent Orthography: An 8-Year Follow-Up  
*Karin Landerl and Heinz Wimmer*

(Contents continue)



- 162 Early First-Language Reading and Spelling Skills Predict Later Second-  
Language Reading and Spelling Skills  
*Richard L. Sparks, Jon Patton, Leonore Ganschow, Nancy Humbach, and  
James Javorsky*
- 175 The Mnemonic Value of Orthography for Vocabulary Learning  
*Julie Rosenthal and Linnea C. Ehri*
- 192 Early Identification of Reading Difficulties Using Heterogeneous  
Developmental Trajectories  
*Christy Kim Boscardin, Bengt Muthén, David J. Francis, and Eva L. Baker*
- 209 The Effects of Tasks on Integrating Information From Multiple Documents  
*Raquel Cerdán and Eduardo Vidal-Abarca*
- 223 A Comparison of Three Measures of Cognitive Load: Evidence for  
Separable Measures of Intrinsic, Extraneous, and Germane Load  
*Krista E. DeLeeuw and Richard E. Mayer*

## Other

- 47 American Psychological Association Subscription Claims Information
- 149 E-Mail Notification of Your Latest Issue Online!
- iii Instructions to Authors
- 161 Low Publication Prices for APA Members and Affiliates
- 66 Members of Underrepresented Groups: Reviewers for Journal Manuscripts  
Wanted
- 77 Subscription Order Form



# Teacher–Student Support, Effortful Engagement, and Achievement: A 3-Year Longitudinal Study

Jan N. Hughes, Wen Luo, Oi-Man Kwok, and Linda K. Loyd  
Texas A&M University

Measures of teacher–student relationship quality (TSRQ), effortful engagement, and achievement in reading and math were collected once each year for 3 consecutive years, beginning when participants were in 1st grade, for a sample of 671 (53.1% male) academically at-risk children attending 1 of 3 school districts in Texas. In separate latent variable structural equation models, the authors tested the hypothesized model, in which Year 2 effortful engagement mediated the association between Year 1 TSRQ and Year 3 reading and math skills. Conduct engagement was entered as a covariate in these analyses to disentangle the effects of effortful engagement and conduct engagement. Reciprocal effects of effortful engagement on TSRQ and of achievement on effortful engagement were also modeled. Results generally supported the hypothesized model. Year 1 variables had a direct effect on Year 3 variables, above year-to-year stability. Findings suggest that achievement, effortful engagement, and TSRQ form part of a dynamic system of influences in the early grades, such that intervening at any point in this nexus may alter children's school trajectories.

**Keywords:** student–teacher relationship, academic engagement, reading, math, elementary grades

Children's academic achievement in the early grades forecasts academic and mental health outcomes throughout their school years and into early adulthood (Alexander, Entwisle, & Horsey, 1997; Campbell, Helms, Sparling, & Ramey, 1998; Entwisle & Alexander, 1988; Finn, 1989; Roeser, Eccles, & Freedman-Doan, 1999; Stevenson & Newman, 1986). Given the importance of a good academic start to school, many researchers have investigated factors that affect children's school readiness skills and early academic trajectories (for reviews, see Future of Children, 2005; Perry & Weinstein, 1998; Shonkoff & Phillips, 2000). Whereas researchers previously focused on child and family contributors to early achievement, recent investigations have assessed the impact of aspects of the classroom and school context that promote or impede children's achievement (Burchinal, Peisner-Feinberg, Pianta, & Howes, 2002; Crosnoe, Johnson, & Elder, 2004; National Institute of Child Health and Human Development Early Child Care Research Network, 2003; Rimm-Kaufman & Pianta, 2000; Roeser et al., 1999). Teacher–student relationship quality (TSRQ) has emerged as an important aspect of the elementary classroom context that has implications for children's concurrent and future academic and social adjustment in school (Birch & Ladd, 1997, 1998; Howes, Hamilton, & Matheson, 1994; Pianta, Steinberg, & Rollins, 1995).

Researchers employing longitudinal designs have found that students who experience teacher–student interactions characterized by high levels of warmth and support or low levels of conflict gain more in achievement (Connell & Wellborn, 1991; Hamre &

Pianta, 2001; Hamre, Pianta, & Downer, 2006; Ladd, Birch, & Buhs, 1999; Pianta & Stuhlman, 2004; Skinner, Zimmer-Gembeck, & Connell, 1998). However, to capitalize on the potential of this classroom resource for improving young children's achievement trajectories, it is important to identify the more proximal mechanisms by which the teacher–student relationship affects achievement. Several investigators have suggested that students who experience an accepting and warm relationship with their teachers will be more capable and motivated to comply with classroom rules and teacher expectations (Brophy, 1983; Furrer & Skinner, 2003; Gest, Welsh, & Domitrovich, 2005; Wentzel, 1998). This increased engagement in classroom learning activities, in turn, is expected to lead to greater achievement gains.

Despite the intuitive appeal of this reasoning, the empirical support for such an indirect model of the effects of TSRQ in the early grades is limited in several ways. First, few studies have employed longitudinal designs that maintain temporal precedence consistent with the hypothesized causal pathways (i.e., TSRQ → engagement → achievement) and statistical controls for prior levels of both engagement and achievement (Skinner et al., 1998). These methodological limitations reduce confidence in conclusions about mediating processes because alternative causal linkages cannot be ruled out (Cole & Maxwell, 2003). Second, the few studies that have been conducted with children in the early school grades have employed measures of classroom engagement that combine different types of engagement (Ladd et al., 1999), thereby making it impossible to disentangle the unique contribution of each type (Fredricks, Blumenfeld, & Paris, 2004). Third, studies have failed to test for the reciprocal effect of engagement on TSRQ. Consistent with transactional (Sameroff, 1975) and developmental systems (Lerner, 1989, 1998) theories, developmental change results from the dynamic interaction between individuals and contexts. Thus, unidirectional models are likely to be inadequate representations of developmental processes.

---

Jan N. Hughes, Wen Luo, Oi-Man Kwok, and Linda K. Loyd, Department of Educational Psychology, Texas A&M University.

Correspondence concerning this article should be addressed to Jan N. Hughes, Texas A&M University, 701 Harrington Building, 4225 TAMU, College Station, TX 77843-4225. E-mail: jhughes@tamu.edu



The purpose of this study was to test an indirect model of the effect of TSRQ on first-grade children's academic achievement over a 3-year period, beginning when children were in first grade. The conceptual model, presented in Figure 1, predicted that TSRQ at Year 1 would affect achievement at Year 3 and that this effect would be mediated by the effect of TSRQ at Year 2. The model included reciprocal causal paths from effortful engagement to TSRQ and from achievement to effortful engagement. To test the unique contribution of effortful engagement to achievement, above the effects of antisocial or conduct engagement, the model also controlled for the effects of prior levels of conduct engagement on Year 3 achievement. Before we review the theoretical and empirical bases for the hypothesized model, we address the multifarious definitions of classroom engagement and the importance of distinguishing between different types of engagement.

### Definitions and Types of Classroom Engagement

Although an extensive literature dating from the 1960s has investigated school and classroom engagement, the construct of engagement has experienced something of a revival in recent years, stimulated by the growing recognition that student disaffection with school increases with additional years in school and is a major factor in student achievement and dropping out of school (Fredricks et al., 2004). In their thorough review of school engagement, Fredricks et al. (2004, p. 60) identified

three broad types of school and classroom engagement: *behavioral engagement* (involvement in academic and social or extracurricular activities), *emotional engagement* (positive and negative reactions to people and activities at school), and *cognitive engagement* (similar to the ideas of investment in learning and intrinsic motivation). Whereas emotional and cognitive engagement have been emphasized in research with middle school and high school students (Connell & Wellborn, 1991; Finn, 1989; Midgley, Feldlaufer, & Eccles, 1989; Ryan, Stiller, & Lynch, 1994; Skinner & Belmont, 1993), behavioral engagement has tended to be the focus of research with elementary students (Alexander, Entwisle, & Dauber, 1993; Birch & Ladd, 1997; Buhs & Ladd, 2001; Miles & Stipek, 2006) and is the focus of the current study. Fredricks et al. (2004) further divided behavioral engagement into three subtypes: conduct, involvement in learning tasks, and participation in extracurricular activities. Studies with elementary students have focused on the first two subtypes of behavior engagement. Conduct engagement is variously defined in terms of antisocial and prosocial behaviors and compliance with classroom rules (Alexander et al., 1993; Gest et al., 2005; Ladd et al., 1999; Miles & Stipek, 2006; Normandeau & Guay, 1998; Trzesniewski, Moffitt, Caspi, Taylor, & Maughan, 2006; Wentzel, 1998). Involvement in learning has been variously defined by "time on task" (Gettinger, 1985; Greenwood, 1991; Rimm-Kaufman, La Paro, Downer, & Pianta, 2005) and by effort, attention, self-direction, and persistence in the classroom (Connell & Well-

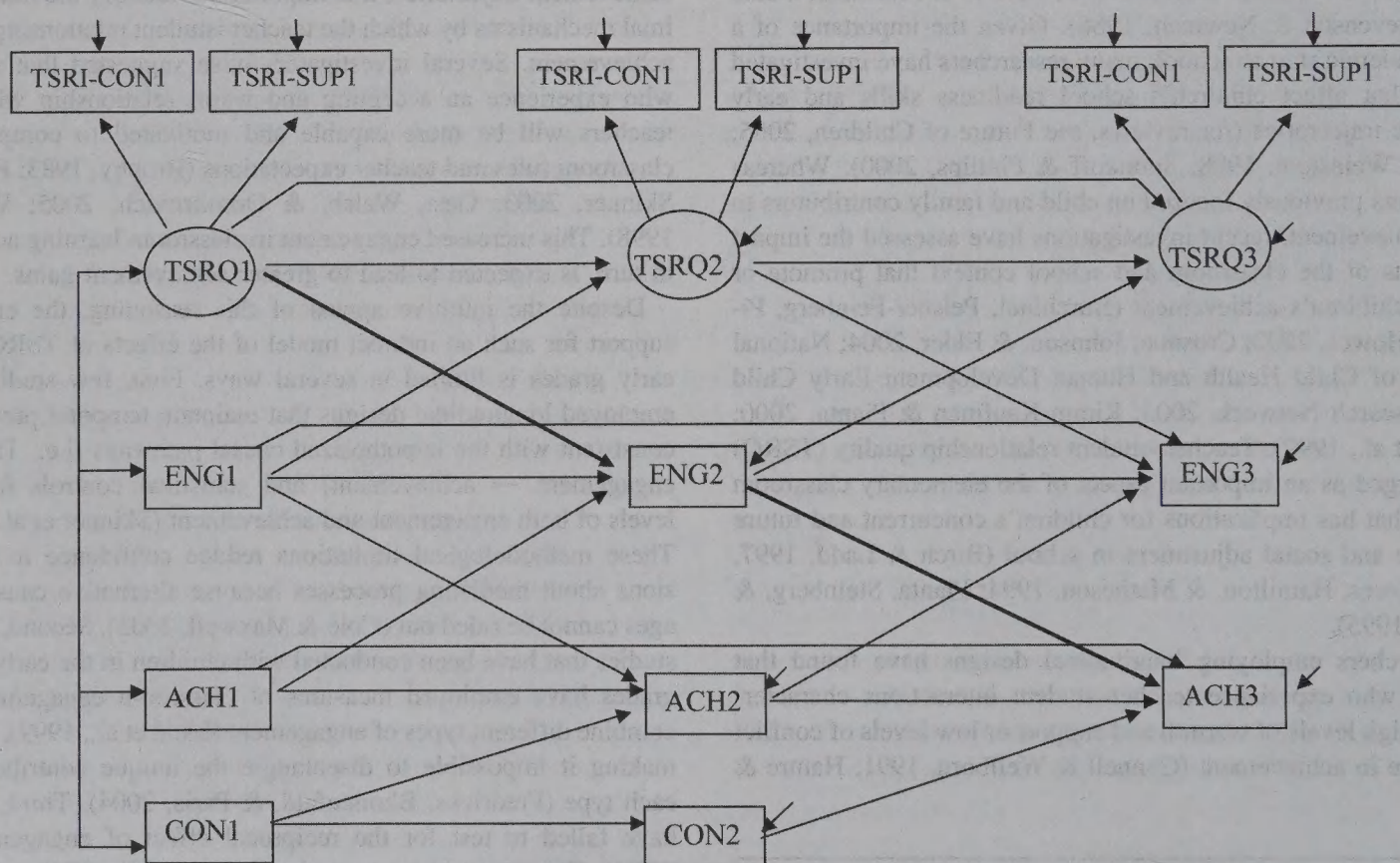


Figure 1. Hypothetical model. The bold lines represent target indirect effects. TSRI-CON = Teacher Student Relationship Inventory Conflict subscale; TSRI-SUP = Teacher Student Relationship Inventory Support subscale; TSRQ = teacher-student relationship quality; ENG = effortful engagement; ACH = either math or reading achievement; CON = conduct engagement.



born, 1991; Furrer & Skinner, 2003; Ladd et al., 1999; Normandeau & Guay, 1998; Skinner & Belmont, 1993).

Because researchers studying the effects of school and classroom engagement on achievement have differed in their definitions and measures of engagement, it is difficult to integrate findings across studies. Often, researchers incorporated a wide variety of constructs in their measurement of engagement, an inclusiveness that makes it difficult to determine the unique precursors and consequences of different types of engagement. In the current study, we assess both conduct engagement and effortful engagement, a construct similar to involvement in learning. *Effortful engagement* refers to the volitional, or effortful, aspect of involvement in instructional activities and includes trying hard, not giving up in the face of difficulty, and directing one's attention to instructional activities. Similar to effortful control (Rothbart & Ahadi, 1994; Rothbart & Bates, 1998; Rothbart & Jones, 1998), effortful engagement is viewed as having a basis in an individual's temperamental impulsivity and attentional capacities yet emerging as a result of the transactions between the child and his or her environment across time. Thus, we expect that effortful engagement will evince moderate stability across time and contexts. Our particular interest is in the effect of the quality of the teacher-student relationship on changes in children's levels of effortful engagement in the classroom and, consequently, on children's academic achievement.

### Effect of Effortful Engagement on Achievement

Longitudinal investigations have documented that high levels of antisocial engagement (or low levels of prosocial behavior) predict declining academic performance (Feldhusen, Thurston, & Benning, 1970; Huesmann, Eron, & Yarmel, 1987; Miles & Stipek, 2006; National Institute of Child Health and Human Development Early Child Care Research Network, 2004; Trzesniewski et al., 2006; Wentzel, 1991). Conversely, high levels of effortful engagement predict improving academic performance (Alexander et al., 1993; Greenwood, 1991; Skinner et al., 1998). However, few researchers have investigated the relative or unique associations between these two types of engagement and achievement (Alexander et al., 1993). Recently, researchers (Miles & Stipek, 2006; Trzesniewski et al., 2006) have speculated that the often found association between antisocial or aggressive conduct and achievement may be due to a "third variable" related both to aggression and achievement, such as cognitive self-control or self-regulated learning behaviors. To the extent that measures of conduct and self-regulated (i.e., effortful) engagement are correlated, the third variable explanation is more plausible. In samples of primary-grade children, Ladd et al. (1999) reported a correlation of .64 between cooperative participation and self-regulated participation, and Normandeau and Guay (1998) reported a correlation of  $-.51$  between aggressive behavior and cognitive self-control, a construct similar to effortful engagement. In their longitudinal study, Normandeau and Guay (1998) found that the effect of conduct engagement in kindergarten on grades 1 year later was accounted for by the effect of conduct engagement on teacher-rated cognitive self-control (e.g., the student sticks to what he or she is doing until finished, persists in face of failure, has to be reminded several times to do something before doing it). On the basis of these findings and the reasoning that effortful engagement in school

learning is more directly related to student mastery of subject matter content than is conduct engagement, we expected that only effortful engagement would uniquely predict subsequent achievement.

### Effect of TSRQ on Effortful Engagement

Researchers have drawn from diverse theoretical conceptualizations in explaining the processes that account for the effect of TSRQ on students' effortful engagement and achievement. On the basis of attachment theory perspectives (Bowlby, 1980), a close and supportive relationship with one's teacher would be expected to promote a child's emotional security and confidence. A secure relationship with the teacher may serve as a resource that permits young students to actively explore their environment and to cope more effectively with novel academic and social demands (Howes et al., 1994; Pianta & Steinberg, 1992). Drawing from this reasoning, Little and Kobak (2003) expected and found that among elementary children, emotional security with the teacher attenuated children's stress reactivity to negative teacher and peer events in the classroom.

Also on the basis of parenting and motivational literature, children who experience warm and close parent-child relationships are more motivated to please their parents and more likely to internalize their parents' values (Dix, 1991; Grusec & Kuczynski, 1997). In a short-term longitudinal study of children in Grades 3-5, teacher support predicted students' liking for school and buffered children with externalizing problems from becoming disaffected with school (Gest et al., 2005). Presumably, children who experience warm and close relationships with their teachers are more likely to identify with school and invest in the school's agenda. Finally, literature on the development of self-regulation (Eisenberg et al., 2005) also supports the view that TSRQ influences children's attention and self-regulation in the classroom. A negative teacher-student relationship might elicit negative emotions in children that interfere with attention and self-regulation (Blair, 2002), whereas a supportive teacher-student relationship might elicit positive moods that promote effective problem solving, regulation, and interactions with others (Isen, Daubman, & Norwicki, 1987; Pianta, 2006).

### Reciprocal Effects Among TSRQ, Effortful Engagement, and Achievement

Consistent with transactional models of development (Sameroff, 1975, 1989), we believe achievement in the early grades is the result of the unfolding of reciprocal processes, such that children's engagement both influences and is influenced by TSRQ and academic competencies. Previous researchers have demonstrated an effect of student antisocial conduct on TSRQ (Birch & Ladd, 1997, 1998; Ladd et al., 1999; Pianta & Steinberg, 1992; Pianta & Stuhlman, 2004). Similarly, it is reasonable to expect that teachers will also find it easier to show support and affection to students who try hard and attend to instructions. It is also likely that children's academic achievement affects, and is affected by, their effortful engagement. For example, Miles and Stipek (2006) found that poor literacy skills in first grade predicted antisocial conduct in third grade. Also supportive of the view that academic performance affects classroom engagement is the finding that academi-



cally focused interventions result in improvements in conduct as well as academic skills (Ayllon & Roberts, 1974; Coie & Krehbiel, 1984).

### Study Hypotheses and Approach

The primary purpose of this study was to test an indirect model of the effects of TSRQ on reading and math achievement via the direct effect of TSRQ on effortful engagement over a 3-year period (beginning when children were in first grade). As Cole and Maxwell (2003) asserted, "a fundamental requirement for one variable to cause another is that the cause must precede the outcome in time" (p. 559). Thus, in testing mediational models with structural equation modeling (SEM), the researcher would ideally collect data on the cause, the mediator, and effect at each of three or more time points (or waves). Such a design not only permits strong assumptions about indirect effects but also permits testing of stationarity and stability of effects, lag effects, and reciprocal causal pathways across children's first 3 years of postkindergarten education.

*Stationarity* implies that the degree to which one set of variables produces change in another set remains the same over time. Perhaps TSRQ in first grade is more important to engagement in Year 2 than TSRQ in Year 2 is to engagement in Year 3. A finding of stronger structural relations between TSRQ and engagement earlier versus later in students' first 3 years of school would be consistent with the view that the quality of children's relationships with their teachers at the beginning of their formal schooling sets in motion patterns of engagement that quickly become self-sustaining. *Stability of effects* refers to the degree to which the within-wave correlations are of the same magnitude at different points in development. *Lag effects* refers to the direct effects of Year 1 variables on Year 3 variables, above the year-to-year stability in the variables. For example, perhaps children's relationships with their first-grade teachers continue to influence their social relatedness with future teachers, above the year-to-year stability in teacher relatedness. Finding such a lag effect for TSRQ would be consistent with the view that children's early relationships with teachers are carried forward as mental representations (Bretherton, 1985) to subsequent relationships with teachers and might explain, in part, the long-term prediction of achievement on the basis of TSRQ in first grade (Pallas, Entwistle, Alexander, & Cadigan, 1987).

The decision to test separate models with reading and math achievement as outcomes instead of using a latent or a composite measure of achievement was based on previous findings that behavior problems predicted reading achievement more strongly than math achievement (Adams, Snowling, Hennessy, & Kind, 1999; DuPaul et al., 2004; Joussemet, Koestner, Lokes, & Landry, 2005). Additionally, because much more time is spent in literacy than in math instruction in early elementary classrooms (National Institute for Child Health and Human Development Early Child Care Research Network, 2003), TSRQ and child engagement might be more predictive of reading than of math achievement.

We also tested whether gender or ethnicity moderated the structural relations in the SEM. Some researchers have found that minority and/or low-socioeconomic status children (Burchinal et al., 2002; Meehan, Hughes, & Cavell, 2003) and boys (Skinner et al., 1998) benefited more from a supportive teacher-student rela-

tionship than did majority children or girls. Other researchers, however, have found no evidence that these demographic variables moderate the relationship between TSRQ and child outcomes (Pianta & Stuhlman, 2004; Silver, Measelle, Armstrong, & Essex, 2005).

We investigated these questions in a culturally and linguistically diverse sample of academically at-risk first-grade children. Children were deemed academically at risk on the basis of scoring below their school district median on a measure of literacy given at the beginning of first grade. Children with relatively low literacy skills in first grade are likely to experience more relational and academic stressors (Ladd et al., 1999; A. J. Reynolds & Bezruczko, 1993). When provided a supportive teacher presence, these children are expected to cope better with stressors and to participate more actively and appropriately in classroom activities, consequently gaining more in achievement. For these reasons, students with low literacy skills represent a population of considerable importance with respect to the effect of teacher support on children's behavioral engagement and academic achievement.

Previously published findings with this longitudinal sample have reported associations between TSRQ, effortful engagement, and achievement using two waves of data (Hughes & Kwok, 2006, 2007). Hughes and Kwok (2006) also reported a mediating role for TSRQ in accounting for the association between students' African American status and effortful engagement. The current study extends these findings by assessing each construct at each of three waves (years) of data, which permits a strong test of mediation, reciprocal effects, and stationarity and stability of effects. Additionally, the current study investigates the unique effects of two types of behavioral engagement, conduct and effortful engagement, on achievement.

## Method

### Participants

Participants were drawn from a larger sample of children participating in a longitudinal study examining the impact of grade retention on academic achievement. Participants were recruited from three school districts in Texas (one urban and two small city) across two sequential cohorts in first grade during the fall of 2001 and 2002. The composition of the urban school was 41% White non-Hispanic, 37% economically disadvantaged, and 11% limited English proficient. Enrollment of one of the small city schools was 40% White non-Hispanic, 61% economically disadvantaged, and 11% limited English proficient. The enrollment of the second small city school was 69% White non-Hispanic, 24% economically disadvantaged, and 5.2% limited English proficient. Children were eligible to participate in the longitudinal study if they scored below the median on a state-approved, district-administered measure of literacy; spoke either English or Spanish; were not receiving special education services; and had not been previously retained in first grade. School records identified 1,374 children as eligible to participate. Because teachers distributed consent forms to parents via children's weekly folders, the exact number of parents who received the consent forms cannot be determined. Incentives, in the form of small gifts to children and the opportunity to win a larger prize in a drawing, were instrumental in



obtaining 1,200 returned consent forms, of which 784 (65%) provided consent.

Analyses on a broad array of archival variables, including performance on the district-administered test of literacy (standardized within district, because of differences in test used), age, gender, ethnicity, eligibility for free or reduced-price lunch, bilingual class placement, cohort, and school context variables (i.e., percentage ethnic/racial minority, percentage economically disadvantaged), did not indicate any difference between children with and without consent. Although we cannot rule out differences between consenters and nonconsenters on variables not included in our data, we can conclude that the resulting sample of 784 participants (52.6% male) closely resembles the eligible sample on demographic and literacy variables relevant to students' educational performance. Of the 784 recruited children, 350 (44.6%) had complete data on all analysis variables assessed at all three occasions (i.e., first grade and 1 and 2 years later), and 671 (85.6%) had data on at least one of the analysis variables at each occasion. Attrition analyses showed that the 350 children with complete data did not differ from the 434 without complete data on demographic variables or study variables at baseline, which supports the assumption that data were missing at random. The overall rate of missingness for the 671 participants with some data at each assessment wave was 12.7%. The 671 participants did not differ from the remaining 113 participants on demographic or study variables at baseline. On the basis of these findings, we imputed the missing value based on these 671 children using SAS PROC MI.

Of these 671 participants, 356 (53.1%) were male, and the racial/ethnic composition was 34.9% White, 36.7% Hispanic, 23.5% African American, and 4.9% Asian/Pacific Islander. At entrance to first grade, children's mean age was 6.57 ( $SD = 0.38$ ) years. Children's mean score for intelligence as measured with the Universal Nonverbal Intelligence Test (McCallum & Bracken, 1997) was 92.89 ( $SD = 14.43$ ). On the basis of family income, 58.0% of participants were eligible for free or reduced-price lunch. For 35.1%, the highest educational level in the household was a high school certificate or less. The ethnic/racial composition for the 337 teachers (94.1% female) completing the teacher questionnaires was 81.3% White, 14.5% Hispanic, 2.7% African American, and 1.5% other ethnicities. The mean teaching experience was 4.42 years ( $SD = 1.82$ ), and 100% of teachers held teacher certification. All teachers had at least a bachelor's degree; 57.6% had done at least some graduate work.

### Design Overview

Assessments were conducted annually for 3 years, beginning when participants were in first grade. Teacher questionnaires assessing teachers' perceptions of the student-teacher relationship and of child engagement were administered in the spring of each year. Teachers received \$25 for completing and returning the questionnaires. Measures of math and reading achievement were individually administered at school at varying times during the school year, with the constraint that at least 8 months separated each annual assessment.

### Measures

**Academic achievement.** The Woodcock-Johnson III (WJ-III) Tests of Achievement (Woodcock, McGrew, & Mather, 2001) is

an individually administered measure of academic achievement for individuals 2 years of age to adulthood. The WJ-III Broad Reading age standard scores (Letter-Word Identification, Reading Fluency, and Passage Comprehension subtests) and the WJ-III Broad Math age standard scores (Calculations, Math Fluency, and Math Calculation Skills subtests) were used. Extensive research documents the reliability and construct validity of scores on the WJ-III and its predecessor (Woodcock & Johnson, 1989; Woodcock et al., 2001).

The Bateria Woodcock-Muñoz: Pruebas de Aprovechamiento—Revisada (Bateria-R; Woodcock & Muñoz-Sandoval, 1996) is the comparable Spanish version of the Woodcock-Johnson Tests of Achievement—Revised (Woodcock & Johnson, 1989), the precursor of the WJ-III. If children or their parents spoke any Spanish, children were administered the Woodcock-Muñoz Language Test (Woodcock & Muñoz-Sandoval, 1993) to determine the child's language proficiency in English and Spanish and selection of either the WJ-III or the Bateria-R. The Woodcock Compuscore (Woodcock & Muñoz-Sandoval, 2001) program yields scores for the Bateria-R that are comparable to scores on the Woodcock-Johnson Tests of Achievement—Revised. The Broad Reading and Broad Mathematics WJ-III scores were used in this study.

**Effortful engagement.** The teacher-report, 10-item effortful engagement scale was composed of 8 items from the Conscientiousness scale of the Big Five Inventory (John & Srivastava, 1999) and 2 items taken from the Social Competence Scale (Conduct Problems Prevention Research Group, 2004) that were consistent with our definition of academic engagement (effort, attention, persistence, and cooperative participation in learning). Although the Big Five Inventory is conceptualized as a measure of personality traits, the selected items from the Conscientiousness scale are similar to items used by other researchers to assess academic engagement (Ladd et al., 1999; Ridley, McWilliam, & Oates, 2000). Items were rated on a 1–5 Likert-type scale. The internal consistency of these 10 items for our sample was .95.

**Conduct engagement.** The teacher-rated measure of conduct (antisocial) engagement was derived from a 24-item questionnaire adapted from the California Child Q-Sort, a language-simplified personality inventory for use by nonprofessionals (Caspi, Block, Block, & Klopp, 1992). Our modification involved use of a rating scale versus Q-sort methodology and a reduction in the number of items from 100 to 24. These 24 items were selected on the basis of previous research demonstrating that they were consistently rated as prototypical of children with high levels of impulsivity and externalizing behaviors (Funder, Block, & Block, 1983). All items were rated on a 1–5 Likert scale.

We applied a cross-validation approach to examine the underlying structure because of the unclear factor structure of the measure (B. Thompson, 2004). The Year 1 data set ( $n = 344$ ) was randomly split into two even halves. We performed an exploratory factor analysis using principal-axis factoring with Promax rotation for the first half of the data; this resulted in a 15-item measure with four factors (9 items were deleted because of high cross-loadings). We then cross-validated the factor structure found from this analysis by performing confirmatory factor analysis on the second half data. Results of the confirmatory factor analysis confirmed the four-factor model,  $\chi^2(112, n = 339) = 225.7, p < .001$  (comparative fit index [CFI] = .97, root-mean-square error of approximation [RMSEA] = .05, standardized root-mean-square residual [SRMR] = .04), and this four-factor model fitted a second sample

adequately,  $\chi^2(112, n = 301) = 241.0, p < .001$  (CFI = .96, RMSEA = .06, SRMR = .05). The four factors were as follows: Prosocial, Antisocial, Ego Resiliency, and Ego Brittle. Model fit was invariant across cohorts and across Waves 1 and 2. Because the Prosocial (4 items;  $\alpha = .93$ ) and Antisocial (4 items;  $\alpha = .86$ ) scales were strongly correlated (standardized path coefficient = .90), we computed a single conduct scale as the mean of the Antisocial and Prosocial (reverse coded) scales. Example Prosocial items included “considerate and thoughtful” and “gets along well with other children.” Example Antisocial items included “physical or verbal aggression” and “tries to take advantage of others.”

*Teacher-reported teacher–student relationship:* The 22-item Teacher Student Relationship Inventory (TSRI; Hughes, Cavell, & Willson, 2001) is based on the Network of Relationships Inventory (Buhrmester & Furman, 1987). Teachers indicate on a 5-point Likert-type scale their level of support (16 items) or conflict (6 items) in their relationships with individual students. An exploratory factor analysis using principal-axis factoring with Promax rotation on 335 first-grade participants from the second cohort suggested three factors: Support (13 items), Intimacy (3 items), and Conflict (6 items). Results of confirmatory factor analysis on 449 first-grade participants from the first cohort found that the three-factor model provided an adequate fit for the data,  $\chi^2(204) = 697.803, p < .001$  (CFI = .92; RMSEA = .074). Furthermore, the null hypothesis of factor invariance across cohorts and times could be retained at the .01 level. Because the Intimacy scale was deemed less relevant to TSRQ, only the Support and Conflict scales were used in the current study. Example Support scale items ( $\alpha = .94$ ) included “I enjoy being with this child,” “This child gives me many opportunities to praise him or her,” and “This child talks to me about things he/she doesn’t want others to know.” Example Conflict scale items ( $\alpha = .91$ ) included “This child and

I often argue or get upset with each other” and “I often need to discipline this child.” In a longitudinal study of behaviorally at-risk elementary students, the TSRI predicted changes in behavioral adjustment and peer relationships (Meehan et al., 2003).

Results

Sample Descriptive Statistics and Intercorrelations

Using multiple imputation, we generated five complete data sets. For simplicity, the sample statistics are reported only for the first data set. Table 1 presents means and standard deviations for analysis variables broken down by gender and ethnicity. Table 2 presents the correlations among the analysis variables. For all the five imputed data sets, we used Green’s (1992) SEM approach to examine the stability of the within-wave correlations across the assessment periods. Green’s method is an SEM-based method to test whether correlation matrices are different from each other for either independent or dependent samples. We tested two models. Model 1 was an unconstrained model in which the within-wave correlations were freely estimated. Model 2 was a constrained model in which the within-wave correlations were constrained to be equal across waves. Then we conducted a chi-square difference test to test whether Model 2 was significantly worse than Model 1. The result was nonsignificant, indicating that we could assume the within-wave correlations were equal across waves. All within-wave correlations were invariant over time for the reading and math models.

Gender and Ethnic Differences

Significant gender differences were found on the measured variables on the basis of the results of one-way multivariate

Table 1  
Means and Standard Deviations of Analysis Variables by Gender and Ethnicity

Scale	Total (N = 671)		Girls (n = 315)		Boys (n = 356)		Majority (Caucasian; n = 234)		Minority (n = 437)	
	M	SD	M	SD	M	SD	M	SD	M	SD
TSRI-CON1	4.10	1.06	4.32	0.91	3.90	1.14	4.18	1.02	4.05	1.08
TSRI-SUP1	4.01	0.80	4.11	0.77	3.79	0.90	4.05	0.78	3.99	0.81
ENG1	3.19	1.09	3.47	1.01	2.95	1.10	3.19	1.05	3.20	1.11
CON1	2.18	0.88	2.05	0.83	2.29	0.90	2.10	0.84	2.22	0.90
READ1	96.36	18.01	98.80	18.15	94.20	17.64	98.61	15.91	95.15	18.95
MATH1	100.62	13.86	100.22	11.96	101.05	13.06	106.89	12.64	97.27	13.32
TSRI-CON2	4.16	1.05	4.40	0.88	3.95	1.13	4.21	0.98	4.14	1.08
TSRI-SUP2	3.92	0.86	4.07	0.79	3.79	0.90	3.95	0.85	3.91	0.87
ENG2	3.35	1.11	3.61	1.06	3.13	1.10	3.46	1.13	3.29	1.09
CON2	0.80	0.96	0.94	0.90	0.67	0.99	0.88	0.92	0.75	0.98
READ2	97.05	16.95	98.95	17.17	95.36	16.60	99.60	14.74	95.68	17.89
MATH2	100.66	12.55	100.22	11.96	101.05	13.06	105.91	12.32	97.85	11.76
TSRI-CON3	4.24	1.00	4.52	0.86	3.99	1.06	4.40	0.91	4.15	1.04
TSRI-SUP3	3.96	0.90	4.11	0.89	3.84	0.89	4.07	0.84	3.91	0.93
ENG3	3.39	0.86	3.61	0.82	3.20	0.84	3.47	0.79	3.35	0.88
READ3	95.30	14.35	97.04	13.52	93.76	14.89	98.43	12.35	93.62	15.06
MATH3	100.83	12.30	100.94	12.29	100.74	12.32	105.38	11.44	98.40	12.05

*Note.* The numbers in the row headings refer to the timing of assessment. TSRI-CON = Teacher Student Relationship Inventory Conflict subscale (teacher perception of conflict); TSRI-SUP = Teacher Student Relationship Inventory Support subscale (teacher perception of warmth); ENG = teacher perception of child academic engagement; CON = conduct engagement; READ = Woodcock–Johnson III Broad Reading age standard score; MATH = Woodcock–Johnson III Broad Math age standard score.



Table 2  
Correlations for All Continuous Analysis Variables

Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1. TSRI-CON1	—																
2. TSRI-SUP1	.60	—															
3. ENG1	.52	.52	—														
4. CON1	.74	.71	.58	—													
5. READ1	.15	.14	.29	.18	—												
6. MATH1	.13	.09	.20	.12	.52	—											
7. TSRI-CON2	.56	.38	.37	.49	.17	.15	—										
8. TSRI-SUP2	.34	.32	.27	.34	.09	.13	.66	—									
9. ENG2	.41	.35	.54	.41	.21	.24	.54	.57	—								
10. CON2	.54	.39	.39	.51	.19	.18	.74	.72	.66	—							
11. READ2	.15	.15	.28	.17	.78	.45	.17	.11	.24	.19	—						
12. MATH2	.17	.13	.27	.17	.48	.75	.17	.19	.28	.23	.54	—					
13. TSRI-CON3	.49	.37	.44	.44	.21	.18	.56	.35	.41	.45	.15	.23	—				
14. TSRI-SUP3	.30	.22	.38	.34	.20	.19	.32	.28	.34	.28	.12	.17	.49	—			
15. ENG3	.37	.30	.52	.37	.25	.23	.35	.28	.50	.34	.22	.24	.50	.58	—		
16. READ3	.16	.13	.32	.19	.73	.43	.15	.10	.28	.17	.83	.51	.18	.14	.29	—	
17. MATH3	.22	.13	.31	.22	.56	.70	.18	.16	.29	.22	.56	.79	.26	.19	.29	.59	—

*Note.* The numbers in the row headings refer to the timing of assessment. TSRI-CON = Teacher Student Relationship Inventory Conflict subscale (teacher perception of conflict); TSRI-SUP = Teacher Student Relationship Inventory Support subscale (teacher perception of warmth); ENG = teacher perception of child academic engagement; CON = conduct engagement; READ = Woodcock-Johnson III Broad Reading age standard score; MATH = Woodcock-Johnson III Broad Math age standard score.

analysis of variance. Over the five imputed data sets, the average  $F(17, 653)$  was 7.01 with a standard deviation of 0.35, and the significance value was less than .01. Girls performed better than boys on TSRI Support: for Year 1,  $F(1, 669) = 13.01$  ( $SD = 2.91$ ),  $p < .001$  ( $ES = .28$ ); for Year 2,  $F(1, 669) = 18.23$  ( $SD = 1.80$ ),  $p < .001$  ( $ES = .33$ ); and for Year 3,  $F(1, 669) = 13.67$  ( $SD = 2.79$ ),  $p < .001$  ( $ES = .28$ ). Girls also outperformed boys on TSRI Conflict: for Year 1,  $F(1, 669) = 33.16$  ( $SD = 5.60$ ),  $p < .001$  ( $ES = .44$ ); for Year 2,  $F(1, 669) = 31.54$  ( $SD = 4.15$ ),  $p < .001$  ( $ES = .43$ ); and for Year 3,  $F(1, 669) = 47.85$  ( $SD = 4.97$ ),  $p < .001$  ( $ES = .53$ ). In addition, girls scored higher than boys on effortful engagement: for Year 1,  $F(1, 669) = 40.36$  ( $SD = 1.68$ ),  $p < .001$  ( $ES = .49$ ); for Year 2,  $F(1, 669) = 29.86$  ( $SD = 2.97$ ),  $p < .001$  ( $ES = .42$ ); and for Year 3,  $F(1, 669) = 38.08$  ( $SD = 3.76$ ),  $p < .001$  ( $ES = .48$ ). Finally, girls outperformed boys on conduct engagement: for Year 1,  $F(1, 669) = 13.73$  ( $SD = 2.35$ ),  $p < .001$  ( $ES = .29$ ); and for Year 2,  $F(1, 669) = 13.87$  ( $SD = 1.17$ ),  $p < .001$  ( $ES = .29$ ). Girls also had higher reading scores at Time 1,  $F(1, 669) = 10.86$  ( $SD = 0.25$ ),  $p < .004$  ( $ES = .22$ ).

Significant racial/ethnic differences were also found on the measured variables. The average  $F(17, 653)$  was 7.49 with a standard error of 0.13, and the significance value was less than .01. Caucasian students performed better on the Time 3 reading test,  $F(1, 669) = 16.82$  ( $SD = 0.72$ ),  $p < .001$  ( $ES = .32$ ), and on math tests at all three time points: for Year 1,  $F(1, 669) = 82.29$  ( $SD = 1.55$ ),  $p < .001$  ( $ES = .70$ ); for Year 2,  $F(1, 669) = 72.99$  ( $SD = 2.93$ ),  $p < .001$  ( $ES = .66$ ); and for Year 3,  $F(1, 669) = 51.92$  ( $SD = 1.13$ ),  $p < .001$  ( $ES = .55$ ).

### Measurement Model for TSRQ

Confirmatory factor analyses were used to examine the factor structure for a teacher-only report of TSRQ ( $N = 671$ ). The latent construct of TSRQ at each occasion was indicated by the TSRI Support and Conflict scales. The items of the Conflict scale were

reverse coded so that a high score indicated low conflict between students and their teachers. To account for the dependency among observations (students) within clusters (classrooms), we conducted all analyses using the “complex analysis” feature in Mplus (Version 3.13; Muthén & Muthén, 2004), in which the models were estimated via the maximum likelihood estimation method with robust standard errors (Muthén & Muthén, 2004). The model fitted the data adequately, with the average  $\chi^2(7) = 32.54$  ( $SD = 8.52$ ),  $p < .01$ , the average CFI = .98 ( $SD = .01$ ), the average RMSEA = .07 ( $SD = .01$ ), and the average SRMR = .03 ( $SD = .004$ ). All the model estimated loadings were significant in a positive direction.

We also examined the invariance of factor loadings of the TSRQ model over time by comparing the chi-square statistics between models with and without constraining the factor loadings of the same indicators to be equal across waves. The difference in chi-square was not significant,  $\Delta\chi^2(2) = 0.36$ ,  $p = .83$ , suggesting that the relations between the two indicators and the latent construct were invariant over time. On the basis of the above time-invariant model, we further examined the factor loading invariance of the TSRQ model between different gender and ethnic groups using the multiple-group comparison approach under the SEM framework. The chi-square difference was significant between different gender groups,  $\Delta\chi^2(1) = 6.21$ ,  $p = .01$ . The average standardized factor loadings of TSRI Conflict and TSRI Support were .90 and .67, respectively, for the girls across waves and .92 and .63, respectively, for the boys across waves. Despite a statistically significant chi-square difference test, the magnitudes of the factor loadings were highly similar for boys and girls, which suggested that the construct of TSRQ was still comparable for girls and boys. There was no significant chi-square difference between different ethnic groups,  $\Delta\chi^2(1) = 0.001$ ,  $p = .98$ , which suggests that the pattern of relations between the two indicators and the latent construct was no different between the majority and minority students.





After taking stationarity into consideration, we obtained a simpler model (i.e., a model with more degrees of freedom because the paths were constrained to be the same across time) that fitted the data equally well, with average  $\chi^2(53) = 239.79$  ( $SD = 12.56$ ),  $p < .001$ , average CFI = .96 ( $SD = .003$ ), average RMSEA = .07 ( $SD = .002$ ), average SRMR = .07 ( $SD = .002$ ). The model with parameter estimates is shown in Figure 2. To provide a clear picture of the major findings, we did not include the covariance estimates between the exogenous variables and the covariance estimates between the residuals within and across waves in Figure 2 but present these estimates in Table 3. In Figure 2, all the unstandardized parameter estimates and the standardized estimates (presented in the parentheses) were the average values over the five imputed data sets. The unstandardized parameter estimates were tested according to the method developed by Schafer (1997). Nonsignificant paths are indicated by the dashed lines in the figure.

To test the effect of Time 1 TSRQ on Time 3 math achievement mediated by Time 2 effortful engagement, we adopted the method developed by MacKinnon, Lockwood, and Williams (2004) to calculate the confidence interval for the mediation effect. If the confidence interval does not include zero, the mediation effect is significant at  $p < .05$ . The point estimate of the mediation effect was .45; the 95% confidence interval of the target mediation effect (i.e., TSRQ at Year 1  $\rightarrow$  engagement at Year 2  $\rightarrow$  math at Year 3; confidence interval = .22, .75) indicated that this mediation effect was significant. The direct effect between TSRQ at Year 1 and math at Year 3 became nonsignificant after we included engagement at Year 2 in the model, which implies that the effect of teacher student-relationship on math achievement was fully me-

diated through effortful engagement. To obtain an effect size estimate of the mediation effect, we calculated the change in squared multiple correlation for the model with and without the two paths that constituted the mediation effect ( $\Delta R^2 = .01$ ). The effects of conduct engagement on math (i.e., conduct engagement at Year 1  $\rightarrow$  math at Year 2; conduct engagement at Year 2  $\rightarrow$  math at Year 3) were not significant.

**Reading achievement.** We repeated the above analyses with the reading achievement model and found a very similar pattern of results. The final model as presented in Figure 3 fitted the data adequately, with average  $\chi^2(52) = 218.46$  ( $SD = 12.50$ ),  $p < .001$ , average CFI = .96 ( $SD = .003$ ), average RMSEA = .07 ( $SD = .003$ ), and average SRMR = .06 ( $SD = .004$ ). Note that covariance estimates between exogenous variables and covariance estimates between the residuals within and across waves are presented in Table 4. The effect of Time 1 TSRQ on Time 3 reading achievement was fully mediated through Time 2 effortful engagement (point estimate = .57; 95% confidence interval of the mediation effect = .27, .93). To obtain an effect size estimate of the mediation effect, we calculated the change in squared multiple correlation for the model with and without the two paths that constituted the mediation effect ( $\Delta R^2 = .10$ ). As was the case for math, the effects of conduct engagement on math (i.e., conduct engagement at Year 1  $\rightarrow$  math at Year 2; conduct engagement at Year 2  $\rightarrow$  math at Year 3) were not significant.

### Moderators

We also examined the possible moderation effects by gender and ethnicity on the transactional relationships among teacher-student relationship, effortful engagement, and achievement. Multiple-group comparison showed that the effects of prior variables on later variables in Figure 2 (i.e., math) and Figure 3 (i.e., reading) were the same for boys and girls: math,  $\Delta\chi^2(15) = 23.74$ ,  $p = .07$ ; reading,  $\Delta\chi^2(16) = 25.09$ ,  $p = .07$ . An examination of the standardized path coefficients confirmed that the structural paths were highly similar for boys and girls. Majority and minority students were also compared in terms of the structural relations, and no differences were found for either math or reading: math,  $\Delta\chi^2(15) = 15.66$ ,  $p = .41$ ; reading,  $\Delta\chi^2(16) = 9.75$ ,  $p = .88$ .

### Discussion

As predicted, the effect of first-grade TSRQ on reading and math achievement 2 years later was completely mediated by Year 2 effortful engagement. This study offers the strongest data yet for the mediating effect of effortful engagement in accounting for the effect of TSRQ on changes in students' achievement. The assessment of TSRQ, effortful engagement, and achievement at each of three time periods provides a strong basis for tests of mediation (Cole & Maxwell, 2003) because the design controlled not only for prior levels of the dependent variable but also for prior levels of the independent variable and the mediator. In the test of the effect of the independent variable on the mediator, prior levels of the mediator were also controlled. Additional strengths are the inclusion of a measure of conduct engagement to determine the unique effect of effortful engagement on achievement and the use of an individually administered measure of reading and math achievement that has strong psychometric properties, rather than teacher

Table 3  
Parameter Estimates of Covariance in the Model Presented in Figure 2

Parameter	Unstandardized estimate	Standardized estimate
Covariance of exogenous variables		
TSRQ1 with ENG1	0.57	.70
TSRQ1 with MATH1	1.66	.16
TSRQ1 with CON1	0.62	.95
ENG1 with MATH1	3.14	.21
ENG1 with CON1	0.56	.58
MATH1 with CON1	1.47	.12
Covariance of correlated residuals		
TSRQ2 with ENG2	0.37	.41
TSRQ2 with MATH2	0.42 <sup>†</sup>	.04
TSRQ2 with CON2	0.48	.61
ENG2 with MATH2	0.45 <sup>†</sup>	.03
ENG2 with CON2	0.40	.39
MATH2 with CON2	0.53 <sup>†</sup>	.05
TSRQ3 with ENG3	0.24	.38
TSRQ3 with MATH3	0.11 <sup>†</sup>	.01
ENG3 with MATH3	0.39 <sup>†</sup>	.04
TSRI-CON1 with TSRI-CON2	0.10	.10
TSRI-CON1 with TSRI-CON3	0.10	.10
TSRI-CON2 with TSRI-CON3	0.14	.14

Note. Estimates with a dagger are not significant at  $p = .05$ . TSRQ = teacher-student relationship quality; ENG = teacher perception of child academic engagement; MATH = Woodcock-Johnson III Broad Math age standard score; CON = conduct engagement; TSRI-CON = Teacher Student Relationship Inventory Conflict subscale (teacher perception of conflict).

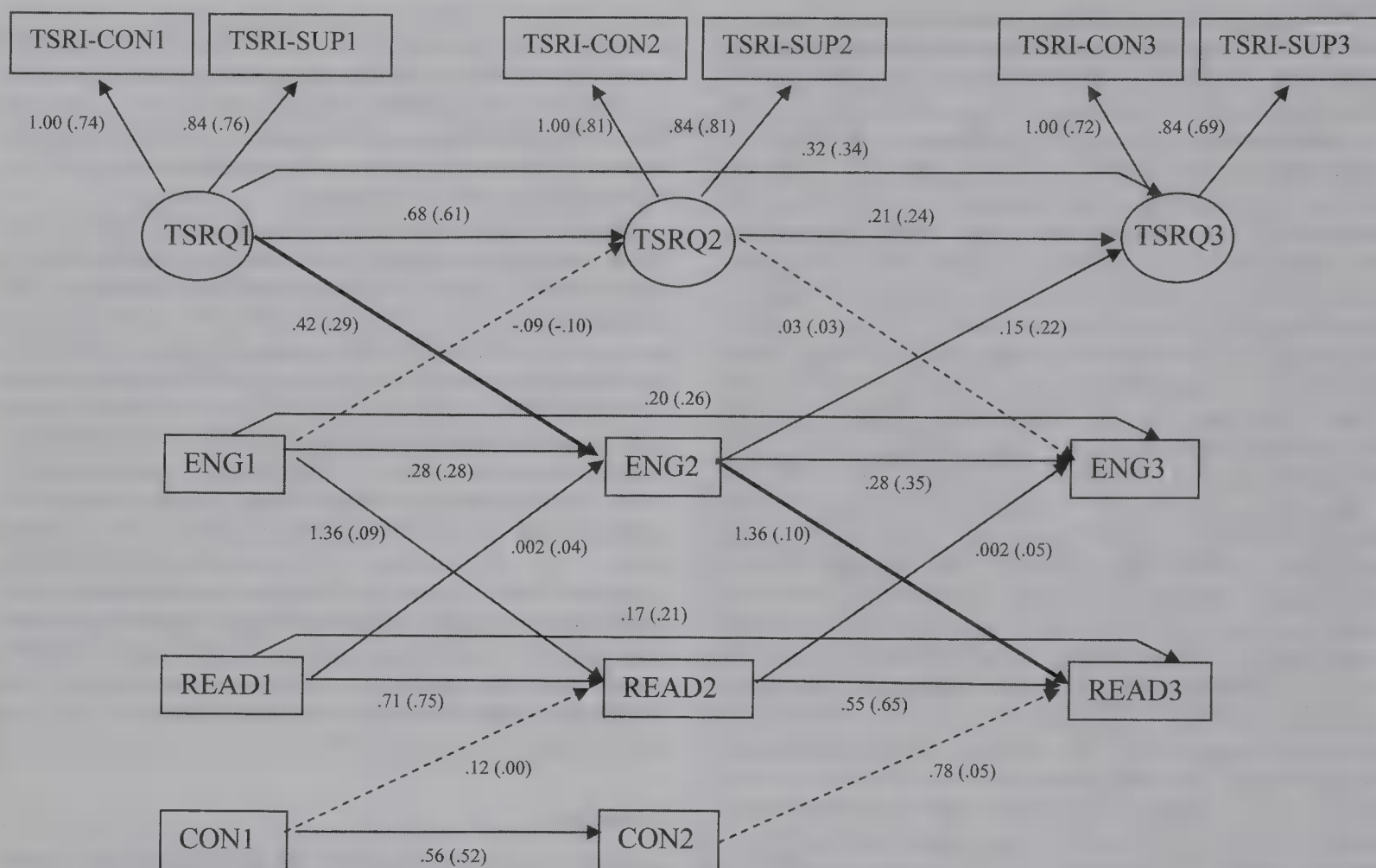


Figure 3. Model of reading achievement. The bold lines represent target indirect effects. The within-wave correlated residuals are not included in the figure for purposes of presentation clarity. Values are unstandardized parameter estimates, with standardized estimates in parentheses. TSRI-CON = Teacher Student Relationship Inventory Conflict subscale; TSRI-SUP = Teacher Student Relationship Inventory Support subscale; TSRQ = teacher-student relationship quality; ENG = effortful engagement; READ = reading performance; CON = conduct engagement.

report of student achievement or grades. The finding that results were nearly identical for models in which the outcome was reading and models in which the outcome was math suggests that the effects were robust across achievement domains. Despite main effects for gender and majority/minority ethnic status, effects were not moderated by gender or ethnicity. Thus, the pattern among these variables and the structural relations were similar for boys and girls and for majority and minority children.

The assessment of study variables at each of three time periods permits strong tests of bidirectional causal pathways. Not only were we able to control for previous levels of each variable in testing each direct effect, we also were able to test these pathways across two time periods. We found support for reciprocal effects between effortful engagement and TSRQ from Year 2 to Year 3 but not from Year 1 to Year 2. This finding might have been due to the higher level of stability for TSRQ from Year 1 to Year 2 than from Year 2 to Year 3.

Reciprocal effects of both math and reading achievement on effortful engagement were invariant across developmental periods. To our knowledge, this is the first study to investigate the effect of achievement on effortful engagement. Children with lower academic skills may become discouraged and believe that their aca-

ademic success is not attributed to their effort. One of the consequences of academic failure is learned helplessness, which is associated with low persistence and effort (Diener & Dweck, 1978). Researchers investigating the effect of achievement on antisocial behavior have speculated that children with low academic skills become frustrated and respond to their frustration with aggression (Miles & Stipek, 2006). Future research is needed to clarify the effect of achievement on different types of engagement and the processes that mediate the effect.

Evidence for stability of effects (e.g., year-to-year stability in measurement of the same construct) was found for effortful engagement and math achievement. However, the effect of Year 1 TSRQ on Year 2 TSRQ was stronger than the effect of Year 2 TSRQ on Year 3 TSRQ. This finding may be a result, in part, of the lag effect of Year 1 TSRQ on Year 3 TSRQ (see the discussion below). Although math demonstrated stability across years, the effect of Year 1 reading on Year 2 reading was stronger than the effect of Year 2 reading on Year 3 reading. Perhaps at older grades factors not included in our model become more important to reading performance.

As predicted, effortful engagement predicted achievement above the effects of prior levels of both conduct engagement and



Table 4  
Parameter Estimates of Covariance in the Model Presented in Figure 3

Parameter	Unstandardized estimate	Standardized estimate
Covariance of exogenous variables		
TSRQ1 with ENG1	0.57	.70
TSRQ1 with READ1	2.54	.19
TSRQ1 with CON1	0.62	.95
ENG1 with READ1	5.74	.29
ENG1 with CON1	0.56	.58
READ1 with CON1	2.77	.18
Covariance of correlated residuals		
TSRQ2 with ENG2	0.37	.41
TSRQ2 with READ2	0.18 <sup>†</sup>	.01
TSRQ2 with CON2	0.48	.61
ENG2 with READ2	0.65 <sup>†</sup>	.04
ENG2 with CON2	0.41	.39
READ2 with CON2	0.08 <sup>†</sup>	.01
TSRQ3 with ENG3	0.24	.39
TSRQ3 with READ3	0.17 <sup>†</sup>	.01
ENG3 with READ3	0.65	.05
TSRI-CON1 with TSRI-CON2	0.10	.10
TSRI-CON1 with TSRI-CON3	0.10	.10
TSRI-CON2 with TSRI-CON3	0.14	.14

*Note.* Estimates with a dagger are not significant at  $p = .05$ . TSRQ = teacher–student relationship quality; ENG = teacher perception of child academic engagement; READ = Woodcock–Johnson III Broad Reading age standard score; CON = conduct engagement; TSRI-CON = Teacher Student Relationship Inventory Conflict subscale (teacher perception of conflict).

achievement. Furthermore, the effect of effortful engagement on achievement was invariant across developmental periods for both reading and math. In contrast, conduct engagement did not contribute to achievement when prior levels of academic engagement and achievement were controlled. To our knowledge, this is the first longitudinal study to investigate the independent effects of these two types of classroom engagement on achievement. Both conduct engagement and effortful engagement probably have their origins in temperament-based self-regulatory competence (Blair, 2002; R. A. Thompson, 1999). Our results suggest that the aspects of self-regulatory competence that affect achievement are only those aspects that interfere with children's ability to attend to instruction and to persevere in academic tasks. These findings suggest that results from studies reporting predictive associations between antisocial behavioral styles and achievement (Miles & Stipek, 2006; Trzesniewski et al., 2006) might be explained by the association between antisocial behavior and constructs similar to our measure of academic engagement. If this is the case, interventions that target student conduct are unlikely to improve academic achievement unless they also improve student academic effort.

This study examined the development of the connections between effortful engagement, achievement, and teacher support in the same children in their first 3 years of formal schooling. Because children's achievement trajectories demonstrate greater stability after third grade (Miles & Stipek, 2006; Skinner et al., 1998), children's patterns of engagement and achievement formed in the first 3 years of formal schooling may have long-lasting impact on their future academic trajectories (Hamre & Pianta, 2001). The lag effect of Year 1 variables (TSRQ, effortful engage-

ment, and achievement) on the corresponding Year 3 variables is consistent with the view that children's early experiences in classrooms launch individual growth trajectories of engagement, achievement, and teacher relatedness. The launch effect may be smaller or disappear as year-to-year changes influence subsequent development. Taken together, these findings suggest that TSRQ in first grade shapes children's patterns of engagement in learning, which leads both to more supportive relationships with subsequent teachers and to higher levels of achievement. Additional years of data collection are needed to determine whether the structural relationships found in the first 3 years of formal schooling persist or whether the associations among teacher support, engagement, and achievement beyond Grade 3 are the result primarily of the stability of each construct.

These results need to be interpreted in the context of study limitations. Because these findings were obtained with a sample of children selected on the basis of scoring below their school district median on a test of early literacy, results may not generalize to children with higher literacy skills. However, the current sample is one of considerable concern to educators and policy makers, given that they scored in the bottom 50% of students in their school districts on a test of literacy. The sample was also ethnically diverse, with an overrepresentation of minority children (65%) relative to the composition of the schools from which these children were selected (58% minority). Thus, study findings point to the potential of interventions in primary grades for reducing racial and ethnic achievement disparities (Hughes & Kwok, 2007).

Another limitation of the study is its reliance on only teacher report for assessment of TSRQ. Initially we had intended to use a multi-informant latent construct of TSRQ based on the TSRI and a peer-nomination measure of teacher–student relationship support (Hughes & Kwok, 2007). Because this measurement model for TSRQ did not fit the data adequately,  $\chi^2(24) = 134.00$ ,  $p < .001$  (CFI = .84; RMSEA = .133; SRMR = .093), we resorted to the latent construct based on the two scales of the TSRI.

Finally, the lack of classroom observational data on teacher–student interactions means the mechanisms that account for the observed effect of TSRQ on students' effortful engagement are not clarified in this study. It may be that teachers who establish positive relationships with students differ in other ways that contribute to students' effort and persistence. For example, teachers who establish positive relationships with students may be more effective in structuring instruction or in managing classroom behavior (Hamre & Pianta, 2005).

As is often the case with longitudinal studies of community samples, missing data were a problem. However, several precautions were taken to ensure that results were not skewed by missing data. Attrition analyses supported the assumption that data were missing at random, and the overall level of missingness was low (12.7%). Furthermore, models estimated with data for those 350 participants who had complete data were nearly identical to results obtained with imputed data for the 671 participants with some data at each of the three waves.

It is well established that children who attend quality preschool programs begin their formal schooling with stronger academic and social skills (for a review, see R. Reynolds, Magnuson, & Ou, 2006). These domains operate in synergistic fashion, such that positive movement in one domain is likely to produce positive movement in other domains. The long-term benefit of such pro-



gram participation may be the result of reciprocal processes between achievement, academic engagement, and TSRQ. Thus, successful interventions at any point in this nexus of influences in the early grades may reverberate in ways that propel positive social and achievement trajectories, and interventions that target all three domains are especially likely to improve achievement.

## References

- Adams, J. W., Snowling, M. J., Hennessy, S. M., & Kind, P. (1999). Problems of behaviour, reading, and arithmetic: Assessments of comorbidity using the Strengths and Difficulties Questionnaire. *British Journal of Educational Psychology*, 69, 571–585.
- Alexander, K. L., Entwisle, D. R., & Dauber, S. L. (1993). First-grade classroom behavior: Its short- and long-term consequences for school performance. *Child Development*, 64, 801–814.
- Alexander, K. L., Entwisle, D. R., & Horsey, C. S. (1997). From first grade forward: Early foundations of high school dropout. *Sociology of Education*, 70, 87–107.
- Ayllon, T., & Roberts, M. (1974). Eliminating discipline problems by strengthening academic performance. *Journal of Applied Behavioral Analysis*, 7, 71–76.
- Bentler, P. M. (2000, September 2). Judea's parameter overidentification. Message posted to SEMNET, archived at <http://bama.ua.edu/cgi-bin/wa?A2=ind0009&L=semnet&P=R1330&I=1>.
- Birch, S. H., & Ladd, G. W. (1997). The teacher-child relationship and children's early school adjustment. *Journal of School Psychology*, 35, 61–79.
- Birch, S. H., & Ladd, G. W. (1998). Children's interpersonal behaviors and the teacher-child relationship. *Developmental Psychology*, 34, 934–946.
- Blair, C. (2002). School readiness: Integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry. *American Psychologist*, 57, 111–127.
- Bowlby, J. (1980). *Attachment and loss*. New York: Basic Books.
- Bretherton, I. (1985). Attachment theory: Retrospect and prospect. *Monographs of the Society for Research in Child Development*, 50(1–2, Serial No. 209), 3–38.
- Brophy, J. (1983). Research on the self-fulfilling prophecy and teacher expectations. *Journal of Educational Psychology*, 75, 631–661.
- Buhrmester, D., & Furman, W. (1987). The development of companionship and intimacy. *Child Development*, 58, 1101–1113.
- Buhs, E. S., & Ladd, G. W. (2001). Peer rejection as an antecedent of young children's school adjustment: An examination of mediating process. *Developmental Psychology*, 37, 550–560.
- Burchinal, M. R., Peisner-Feinberg, E., Pianta, R., & Howes, C. (2002). Development of academic skills from preschool through second grade: Family and classroom predictors of developmental trajectories. *Journal of School Psychology*, 40, 415–436.
- Campbell, F. A., Helms, R., Sparling, J. J., & Ramey, C. T. (1998). Early childhood programs and success in school: The Abecedarian study. In W. S. Barnett & S. S. Boocock (Eds.), *Early care and education for children in poverty: Promises, programs and long-term results* (pp. 145–166). New York: State University of New York Press.
- Caspi, A., Block, J., Block, J. H., & Klopp, B. (1992). A "common-language" version of the California Child Q-Set for Personality Assessment. *Psychological Assessment*, 4, 512–523.
- Coie, J. D., & Krehbiel, G. (1984). Effects of academic tutoring on the social status of low-achieving, socially rejected children. *Child Development*, 55, 1465–1478.
- Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, 112, 558–577.
- Conduct Problems Prevention Research Group. (2004). *Teacher social competence*. Retrieved September 24, 2004, from <http://www.fasttrackproject.org/techrept/t/tsc/>
- Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy, and relatedness: A motivational analysis of self-system processes. In M. Gunnar & L. A. Sroufe (Eds.), *Minnesota symposium on child psychology* (Vol. 22, pp. 43–77). Hillsdale, NJ: Erlbaum.
- Crosnoe, R., Johnson, M. K., & Elder, G. H. (2004). Intergenerational bonding in school: The behavioral and contextual correlates of student-teacher relationships. *Sociology of Education*, 77, 60–81.
- Diener, C. I., & Dweck, C. S. (1978). An analysis of learned helplessness: Continuous changes in performance, strategy, and achievement cognitions following failure. *Journal of Personality and Social Psychology*, 36, 451–462.
- Dix, T. (1991). The affective organization of parenting: Adaptive and maladaptive processes. *Psychological Bulletin*, 110, 3–25.
- DuPaul, G. J., Volpe, R. J., Jitendra, A. K., Lutz, J. G., Lora, K. S., & Gruber, R. (2004). Elementary school students with AD/HD: Predictors of academic achievement. *Journal of School Psychology*, 42, 285–301.
- Eisenberg, N., Zhou, Q., Spinrad, T. L., Valiente, C., Fabes, R. A., & Liew, J. (2005). Relations among positive parenting, children's effortful control, and externalizing problems: A three-wave longitudinal study. *Child Development*, 76, 1055–1071.
- Entwisle, D. R., & Alexander, K. L. (1988). Factors affecting achievement test scores and marks of Black and White first graders. *Elementary School Journal*, 88, 449–471.
- Feldhusen, J. F., Thurston, J. R., & Benning, J. J. (1970). Longitudinal analyses of classroom behavior and school achievement. *Journal of Experimental Education*, 38, 4–10.
- Finn, J. D. (1989). Withdrawing from school. *Review of Educational Research*, 59, 117–142.
- Fredericks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74, 59–109.
- Funder, D. C., Block, J. H., & Block, J. (1983). Delay of gratification: Some longitudinal personality correlates. *Journal of Personality and Social Psychology*, 44, 1198–1213.
- Furrer, C., & Skinner, E. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology*, 95, 148–162.
- Future of Children. (2005). School readiness: Closing racial and ethnic gaps. The Future of Children, Vol. 15. The Woodrow Wilson School of Public and International Affairs at Princeton University and the Brookings Institute. Retrieved February 15, 2005 from [http://www.futureofchildren.org/pubs-info2825/pubs-info\\_show.htm?doc\\_id=255946](http://www.futureofchildren.org/pubs-info2825/pubs-info_show.htm?doc_id=255946)
- Gest, S. D., Welsh, J. A., & Domitrovich, C. E. (2005). Behavioral predictors of changes in social relatedness and liking school in elementary school. *Journal of School Psychology*, 43, 281–301.
- Gettlinger, M. (1985). Time allocated and time spent relative to time needed for learning as determinants of achievement. *Journal of Educational Psychology*, 77, 3–11.
- Green, J. A. (1992). Testing whether correlation matrices are different from each other. *Developmental Psychology*, 28, 215–224.
- Greenwood, C. R. (1991). Longitudinal analysis of time, engagement, and achievement in at-risk versus non-risk students. *Exceptional Children*, 57, 521–536.
- Grusec, J. E., & Kuczynski, L. (1997). *Parenting and children's internalization of values*. New York: Wiley.
- Hamre, B. K., & Pianta, R. C. (2001). Early teacher-child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development*, 72, 625–638.
- Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development*, 76, 949–967.



- Hamre, B. K., Pianta, R. C., & Downer, J. (2006, May). *Social resources in the classroom and young children's academic and social development*. Paper presented at the meeting of the Society for Prevention Research, San Antonio, TX.
- Howes, C., Hamilton, C. E., & Matheson, C. C. (1994). Children's relationships with peers: Differential associations with aspects of the teacher-child relationship. *Child Development, 65*, 253-263.
- Huesmann, L. R., Eron, L. D., & Yarmel, P. W. (1987). Intellectual functioning and aggression. *Journal of Personality and Social Psychology, 52*, 232-240.
- Hughes, J. N., Cavell, T. A., & Willson, V. (2001). Further support for the developmental significance of the quality of the teacher-student relationship. *Journal of School Psychology, 39*, 289-301.
- Hughes, J. N., & Kwok, O. (2006). Classroom engagement mediates the effect of teacher-student support on elementary students' peer acceptance: A prospective analysis. *Journal of School Psychology, 43*, 465-480.
- Hughes, J. N., & Kwok, O. (2007). The influence of student-teacher and parent-teacher relationships on lower achieving readers' engagement and achievement in the primary grades. *Journal of Educational Psychology, 99*, 39-51.
- Isen, A. M., Daubman, K. A., & Norwicki, G. P. (1987). The influence of affect on categorization. *Journal of Personality and Social Psychology, 47*, 1206-1217.
- John, O. P., & Srivastava, S. (1999). *The Big Five trait taxonomy: History, measurement, and theoretical perspectives*. New York: Guilford Press.
- Joussemet, M., Koestner, R., Lekes, N., & Landry, R. (2005). A longitudinal study of the relationship of maternal autonomy support to children's adjustment and achievement in school. *Journal of Personality, 73*, 1215-1235.
- Ladd, G. W., Birch, S. H., & Buhs, E. S. (1999). Children's social and scholastic lives in kindergarten: Related spheres of influence? *Child Development, 70*, 1373-1400.
- Lerner, R. M. (1989). Developmental contextualism and the life-span view of person-context interaction. In M. H. Bornstein & J. S. Bruner (Eds.), *Interaction in human development: Crosscurrents in contemporary psychology* (pp. 217-239). Hillsdale, NJ: Erlbaum.
- Lerner, R. M. (1998). Theories of human development: Contemporary perspectives. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology* (Vol. 1, 5th ed., pp. 1-24). New York: Wiley.
- Little, M., & Kobak, R. (2003). Emotional security with teachers and children's stress reactivity: A comparison of special-education and regular-education classrooms. *Journal of Clinical Child and Adolescent Psychology, 32*, 127-138.
- MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research, 39*, 99-128.
- McCallum, R. S., & Bracken, B. A. (1997). The Universal Nonverbal Intelligence Test. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 268-280). New York: Guilford Press.
- Meehan, B. T., Hughes, J. N., & Cavell, T. A. (2003). Teacher-student relationships as compensatory resources for aggressive children. *Child Development, 74*, 1145-1157.
- Midgley, C., Feldlaufer, H., & Eccles, J. (1989). Student/teacher relations and attitudes toward mathematics before and after the transition to junior high school. *Child Development, 60*, 981-992.
- Miles, S. B., & Stipek, D. (2006). Contemporaneous and longitudinal associations between social behavior and literacy achievement in a sample of low-income elementary school children. *Child Development, 77*, 103-117.
- Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide*. Los Angeles: Muthén & Muthén.
- National Institute of Child Health and Human Development Early Child Care Research Network. (2003). Social functioning in first grade: Associations with earlier home and child care predictors and with current classroom experiences. *Child Development, 74*, 1639-1662.
- National Institute of Child Health and Human Development Early Child Care Research Network. (2004). Trajectories of physical aggression from toddlerhood to middle childhood. *Monographs of the Society for Research in Child Development, 69*(4, Serial No. 2780).
- Normandeau, S., & Guay, F. (1998). Preschool behavior and first-grade school achievement: The mediational role of cognitive self-control. *Journal of Educational Psychology, 90*, 111-121.
- Pallas, A. M., Entwisle, D. R., Alexander, K. L., & Cadigan, D. (1987). Children who do exceptionally well in first grade. *Sociology of Education, 60*, 257-271.
- Perry, K. E., & Weinstein, R. S. (1998). The social context of early schooling and children's school adjustment. *Educational Psychologist, 33*, 177-194.
- Pianta, R. C. (2006). Teacher-child relationships and early literacy. In D. Dickinson & S. Neuman (Eds.), *Handbook of early literacy research* (Vol. 2, pp. 149-162). New York: Guilford Press.
- Pianta, R. C., & Steinberg, M. S. (1992). Teacher-child relationships and the process of adjusting to school. *New Directions for Child Development, 57*, 61-80.
- Pianta, R. C., Steinberg, M. S., & Rollins, K. B. (1995). The first two years of school: Teacher-child relationships and deflections in children's classroom adjustment. *Development and Psychopathology, 7*, 295-312.
- Pianta, R. C., & Stuhlman, M. W. (2004). Teacher-child relationships and children's success in the first years of school. *School Psychology Review, 33*, 444-458.
- Reynolds, A. J., & Bezruczko, N. (1993). School adjustment of children at risk through fourth grade. *Merrill-Palmer Quarterly, 39*, 457-480.
- Reynolds, R., Magnuson, K., & Ou, S. R. (2006). *PK-3 education: Programs and practices that work in children's first decade*. Retrieved April 23, 2006, from the Foundation for Child Development Web site: <http://www.fcd-us.org/pdfs/PK-3EducationProgramsandPracticesthatWork.pdf>
- Ridley, S. M., McWilliam, R. A., & Oates, C. S. (2000). Observed engagement as an indicator of child care program quality. *Early Education and Development, 11*, 133-146.
- Rimm-Kaufman, S. E., La Paro, K. M., Downer, J. T., & Pianta, R. C. (2005). The contribution of classroom setting and quality of instruction to children's behavior in kindergarten classrooms. *Elementary School Journal, 105*, 377-394.
- Rimm-Kaufman, S. E., & Pianta, R. C. (2000). An ecological perspective on the transition to kindergarten: A theoretical framework to guide empirical research. *Journal of Applied Developmental Psychology, 21*, 491-511.
- Roeser, R. W., Eccles, J. S., & Freedman-Doan, C. (1999). Academic functioning and mental health in adolescence: Patterns, progressions, and routes from childhood. *Journal of Adolescent Research, 14*, 135-174.
- Rothbart, M. K., & Ahadi, S. (1994). Temperament and the development of personality. *Journal of Abnormal Psychology, 103*, 55-66.
- Rothbart, M. K., & Bates, J. E. (1998). Temperament. In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (5th ed., pp. 105-176). New York: Wiley.
- Rothbart, M. K., & Jones, L. (1998). Temperament, self-regulation, and education. *School Psychology Review, 27*, 479-491.
- Ryan, R. M., Stiller, J. D., & Lynch, J. H. (1994). Representations of relationships to teachers, parents, and friends as predictors of academic motivation and self-esteem. *Journal of Early Adolescence, 14*, 226-249.
- Sameroff, A. J. (1975). Transactional models in early social relations. *Human Development, 18*, 65-79.
- Sameroff, A. J. (1989). Principles of development and psychopathology. In



- A. Sameroff & R. Emde (Eds.), *Relationship disturbances in early childhood* (pp. 17–32). New York: Basic Books.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
- Shonkoff, J. P., & Phillips, D. (Eds.). (2000). *From neurons to neighborhoods: The science of early childhood development*. Washington, DC: National Academy Press.
- Silver, R. B., Measelle, J. R., Armstrong, J. M., & Essex, M. J. (2005). Trajectories of classroom externalizing behavior: Contributions of child characteristics, family characteristics, and the teacher-child relationship during the school transition. *Journal of School Psychology, 43*, 39–60.
- Skinner, E., & Belmont, M. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology, 85*, 571–581.
- Skinner, E. A., Zimmer-Gembeck, M. J., & Connell, J. P. (1998). Individual differences and the development of perceived control. *Monographs of the Society for Research in Child Development, 63*(2–3, Serial No. 254).
- Stevenson, H. W., & Newman, R. S. (1986). Long-term prediction of achievement and attitudes in mathematics and reading. *Child Development, 57*, 646–659.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.
- Thompson, R. A. (1999). The individual child: Temperament, emotion, self, and personality. In M. H. Bornstein & M. E. Lamb (Eds.), *Developmental psychology: An advanced textbook* (4th ed., pp. 377–410). Mahwah, NJ: Erlbaum.
- Trzesniewski, K. H., Moffitt, T. E., Caspi, A., Taylor, A., & Maughan, B. (2006). Revisiting the association between reading achievement and antisocial behavior: New evidence of an environmental explanation from a twin study. *Child Development, 77*, 72–88.
- Wentzel, K. R. (1991). Social competence at school: Relation between social responsibility and academic achievement. *Review of Educational Research, 61*, 1–24.
- Wentzel, K. (1998). Social relationships and motivation in middle school: The role of parents, teachers, and peers. *Journal of Educational Psychology, 90*, 202–209.
- Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-Educational Battery—Revised*. Allen, TX: DLM Teaching Resources.
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement*. Itasca, IL: Riverside Publishing.
- Woodcock, R. W., & Muñoz-Sandoval, A. F. (1993). *Woodcock-Muñoz Language Survey*. Chicago: Riverside.
- Woodcock, R. W., & Muñoz-Sandoval, A. F. (1996). *Batería Woodcock-Muñoz: Pruebas de Aprovechamiento—Revisada*. Itasca, IL: Riverside Publishing.
- Woodcock, R. W., & Muñoz-Sandoval, A. F. (2001). *Woodcock-Muñoz Language Survey Normative Update scoring and reporting program* [Computer software]. Itasca, IL: Riverside Publishing.

Received September 12, 2006

Revision received June 21, 2007

Accepted June 28, 2007 ■



# Interplay Between Personal Goals and Classroom Goal Structures in Predicting Student Outcomes: A Multilevel Analysis of Person–Context Interactions

Shun Lau and Youyan Nie  
Nanyang Technological University

This study examined cross-level interactions between personal goals and classroom goal structures, as well as their additive contributions to predicting math achievement, engagement, interest, effort withdrawal, and avoidance coping, using a sample of 3,943 Grade 5 students from 130 classrooms. Results of hierarchical linear modeling showed that classroom performance goal structures exacerbated (a) the negative association between personal performance-avoidance goals and engagement and (b) the positive relations of personal performance-avoidance goals to effort withdrawal and avoidance coping. Moreover, both classroom performance goal structures and personal performance-avoidance goals had maladaptive patterns of relations to outcomes at their respective levels of analysis, whereas classroom mastery goal structures and personal mastery goals showed adaptive relations. Our findings underscore the importance of a multilevel interactionist perspective in understanding achievement motivation and making recommendations for educational practices.

**Keywords:** classroom goal structure, personal goal orientation, person–context interaction, achievement goal theory, elementary school

In the achievement goal literature, two lines of research have received considerable attention. The first line of research reflects the *person* perspective of motivation and focuses on investigating the motivational dynamics of individuals who adopt different types of personal goals (Dweck, 1986; Dweck & Leggett, 1988; Elliot, 2005). The second line of research reflects the *contextualist* perspective and focuses on how different types of contextual goal structures (salient goal-related messages conveyed by classroom practices or school policies) influence achievement-related behavior in educational settings (Ames, 1992; Ames & Archer, 1988; Maehr & Midgley, 1996). Although the two perspectives are complementary and mutually informative (Roeser, 2004), relatively little research has integrated the person and contextualist perspectives to study motivational phenomena in achievement contexts (Linnenbrink, 2004, 2005). In this study, we examined the joint (additive and interactive) contributions of classroom goal

structures and personal goals to the prediction of students' achievement and motivational outcomes from a *multilevel interactionist* perspective. A specific emphasis of this article is a clear distinction between the personal and the contextual levels in our conceptualization and statistical analysis of person–context interactions (Chan, 2006; Glick, 1985; Raudenbush & Bryk, 2002).

## Person Perspective

For researchers who are oriented toward the person perspective, one of their primary objectives is to build a science of achievement motivation with achievement goals as the central organizing constructs (Dweck, 1986; Dweck & Leggett, 1988; Elliot, 2005; Pintrich, 2000). In educational settings, students' achievement goals represent their reasons or purposes for engaging in academic tasks. A large body of empirical evidence has been accumulated in relation to three types of personal goals:<sup>1</sup> (a) personal mastery goals, which emphasize learning new knowledge and improving one's skills; (b) personal performance-approach goals, which emphasize demonstrating one's competence relative to others; and (c) personal performance-avoidance goals, which emphasize the avoidance of showing one's incompetence relative to others. The general findings are that personal mastery goals and performance-avoidance goals are consistently associated with adaptive and maladaptive outcomes, respectively. However, empirical evidence

---

Shun Lau and Youyan Nie, Centre for Research in Pedagogy and Practice, National Institute of Education, Nanyang Technological University, Singapore.

We thank David Hogan, Kin-Meng Lim, Ridzuan Abdul Rahim, Yee-Zher Sheng, Gim-Hoon Tan, and Louis Tay for their assistance in this research. We also thank Robert Roeser for his conceptual insights.

This research was supported by the Centre for Research in Pedagogy and Practice under a Singapore Ministry of Education research grant. Any opinions, findings, and conclusions expressed in this article are those of the authors and do not reflect the views of the Centre for Research in Pedagogy and Practice or the Singapore Ministry of Education.

Correspondence concerning this article should be addressed to Shun Lau, Centre for Research in Pedagogy and Practice, National Institute of Education, Nanyang Technological University, 1 Nanyang, Walk, Singapore 637616. Email: shun.lau@nie.edu.sg

---

<sup>1</sup> In the latest formulation of achievement goal constructs, mastery goals were bifurcated into mastery-approach and mastery-avoidance goals (Elliot & McGregor, 2001; Pintrich, 2000). In this article, we focused on the approach dimension of the mastery goal and both the approach and avoidance dimensions of the performance goal because these three personal goals have produced the most solid empirical base.



for the role of performance-approach goals in achievement-related outcomes is mixed (see Elliot, 2005; Harackiewicz, Barron, Pintrich, Elliot, & Thrash, 2002; Midgley, Kaplan, & Middleton, 2001; Pintrich & Schunk, 2002, for recent reviews).

### Contextualist Perspective

In a parallel development, practice-oriented and reform-minded researchers who are oriented toward the contextualist perspective have turned their attention to the influences of contextual (classroom or school) goal structures on students' achievement motivation (Ames, 1992; Ames & Archer, 1988; Anderman, Maehr, & Midgley, 1999; Maehr & Midgley, 1996; Roeser, 2004). One of their primary concerns is to identify contextual factors that are systematically related to key educational outcomes. Much research has focused on two types of classroom goal structures derived from the achievement goal theory: (a) the classroom mastery goal structure and (b) the classroom performance goal structure.<sup>2</sup> In mastery-oriented classrooms, instructional practices, task assignment, and evaluation procedure are structured to emphasize learning, task mastery, and trying hard to improve one's skills. In contrast, in performance-oriented classrooms, instructional practices, task assignment, and evaluation procedure are structured to emphasize demonstrating competence. The general findings are that classroom mastery goal structures are related to adaptive outcomes and classroom performance goal structures are related to maladaptive outcomes, although variations exist depending on the specific outcomes under investigation (see Kaplan, Middleton, Urdan, & Midgley, 2002; Meece, Anderman, & Anderman, 2006; Urdan, 2004, for recent reviews).

### Level of Analysis Problem

Although research based on the person and contextualist perspectives has yielded different types of information (the former about motivational dynamics of individuals with different profiles of personal goals; the latter about what classroom or school environments would enhance or impede students' learning and motivation), considerable confusion has arisen from attempts to make substantive conclusions about influences of contextual goal structures by using evidence derived from the personal level of analysis (Roeser, 2004). Such confusion is exemplified by a recent debate that centered on the "unique positive potential of performance-approach goals, relative to mastery goals" (Harackiewicz, et al., 2002, p. 638; Midgley et al., 2001). Although the evidence that was marshaled to support each side of the debate was primarily based on the individual level of analysis, both Midgley et al. and Harackiewicz et al. discussed the broader educational implications of the evidence at the contextual level. For example, Midgley et al. were concerned about the unsettling implication that proving the adaptiveness of personal performance-approach goals might convey a message to teachers that they should emphasize performance goals in their classrooms—that "[t]he call for a reconceptualization of goal theory gives the message that it may be alright to emphasize the demonstration of ability and competition among students" (p. 83). In contrast, Harackiewicz et al. argued that emphasizing multiple goals in the classroom would provide students with multiple pathways to achieve the same valued outcomes—that "there may be multiple pathways to improve schools, not just one 'mas-

tery road' that all must travel" (p. 643). Thus, there seems to be an implicit assumption that what is true at the individual level of analysis is also true at the contextual level of analysis (Roeser, 2004).

In this article, we argue that person-level findings cannot be directly translated into conclusions at the contextual level and that direct evidence derived from the contextual level of analysis is required to appropriately inform classroom practices or school policies. Although evidence has been reported to show either the direct relations between contextual goal structures and student outcomes or the indirect relations mediated by personal goals (e.g., Church, Elliot, & Gable, 2001; Kaplan & Midgley, 1999; Nolen & Haladyna, 1990; Roeser, Midgley, & Urdan, 1996), the reliance on single-level analyses in many of the reported studies has created interpretive difficulty.

Conceptually, a variable takes on different meanings at different levels of analysis (Chan, 2006; Glick, 1985; Raudenbush & Bryk, 2002). In relation to achievement goal research, if a researcher's objective is to study the contextual effects of classroom goal structures on student outcomes, the level of conceptualization is the classroom and not the individual. Accordingly, in statistical analysis, classroom goal structures should be treated as level-2 variables. In contrast, if a researcher's objective is to study the psychological correlates of personal goals, the level of conceptualization is the individual and not the classroom. Accordingly, in statistical analysis, personal goals should be treated as level-1 variables, and the conclusion drawn from this level of analysis should only be applicable to level 1. Statistically, ignoring the nested structure of the data and reliance on single-level analysis can lead to misleading estimates of parameters (Cronbach, 1976; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999).

### Are Person-Level Findings Consistent With Classroom-Level Findings?

Recent methodological advances in multilevel analysis or hierarchical linear modeling (HLM) have allowed researchers to analyze hierarchical data in a multilevel framework, thus resolving the level of analysis problem (Goldstein, 2003; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). The most relevant and informative studies are those that included both classroom goal structures and personal goals in the same HLM analysis. These studies have enabled direct comparisons between the role of classroom goal structures and personal goals with the same sample, measurement, and design characteristics. Moreover, they have provided evidence about the unique contribution of classroom goal structures after accounting for the predictive contribution of personal goals. This type of research, however, is limited in number and has yielded mixed findings. For example, Karabenick (2004) found that, in college introductory chemistry classrooms, level-1 (person-level) relations showed differential consistency with level-2 (classroom-level) relations, depending on the outcome variables used in the

<sup>2</sup> Some recent research has distinguished between the classroom performance-approach and classroom performance-avoidance goal structures (e.g., Kaplan, Gheen, & Midgley, 2002; Karabenick, 2004; Wolters, 2004), but the empirical support for such a distinction is not well established (see Discussion section for an elaboration of this point).



analysis. Specifically, at level 1, personal mastery and performance-approach goals were significant positive predictors of help-seeking approach, whereas personal performance-avoidance goals were significant negative predictors. At level 2, however, classroom mastery, performance-approach, and performance-avoidance goal structures were not significant predictors of help-seeking approach. When help-seeking avoidance was used as the outcome variable, level-1 relations were consistent with level-2 relations. Personal performance-avoidance goals were significant positive predictors at level 1 and classroom performance-avoidance goal structures were significant positive predictors at level 2. The other two personal goals and classroom goal structures were not significant predictors at level 1 and level 2, respectively. In another HLM study, Kaplan, Gheen, and Midgley (2002) found that level-1 and level-2 relations were quite consistent in a sample of Grade 9 students in math classrooms. Specifically, at level 1, personal mastery goals were negative predictors of disruptive behavior, whereas personal performance-approach and performance-avoidance goals were positive predictors. At level 2, classroom mastery goal structures were negative predictors of disruptive behavior, whereas classroom performance-approach goal structures were positive predictors.

Although these studies have significantly contributed to our understanding of multilevel relations, the paucity of research that included both classroom goal structures and personal goals in the same HLM analysis as well as the inconsistency of evidence has limited the generalizability of the findings to different outcomes and different age groups. Accordingly, one of the primary objectives of this study was to examine the predictive relations of classroom goal structures and personal goals to an expanded array of outcomes in Grade 5 math classrooms, including academic achievement, interest, engagement, effort withdrawal, and avoidance coping. The inclusion of both achievement and interest as outcomes is particularly pertinent in light of Harackiewicz et al.'s (2002) review of evidence in support of the differential relations that personal mastery goals were positively related to interest, but not achievement, whereas personal performance-approach goals were positively related to achievement, but not interest. Moreover, the inclusion of the three other outcome variables would allow us to examine the role of classroom goal structures and personal goals in both adaptive and maladaptive functioning of students. It is our aim that this study would capture the best features of the person and contextualist perspectives by advancing our knowledge of motivational dynamics at the student level as well as informing educational practices at the classroom level.

### From Additive Models to Interactive Models

In the preceding sections, we have discussed the rationale and evidence in support of moving from single-level analysis to multilevel analysis. The HLM research we have reviewed employed additive multilevel models to explore the unique contributions of classroom goal structures and person goals to predicting outcomes at different levels. Additive multilevel models assume that relations between level-1 variables (e.g., between personal goals and student-level outcomes) are homogenous across different contexts. This assumption needs to be tested explicitly in order to determine whether level-1 relations vary across classroom contexts and, if so, how classroom contexts may moderate (strengthen or weaken) the

relations at level 1. The multilevel interactionist framework would be most suitable for addressing these questions.

### Multilevel Interactionist Perspective

Consistent with the trend toward interactive models in developmental psychology (e.g., Bergman, Cairns, Nilsson, & Nystedt, 2000; Magnusson & Stattin, 1998), personality-social psychology (e.g., Mischel, 2003; Mischel & Shoda, 1995), educational psychology (e.g., Linnenbrink, 2004; Linnenbrink & Pintrich, 2001), and psychopathology (e.g., Ingram & Luxton, 2005; Rutter et al., 1997), we conceptualized students' achievement motivation in the classroom from an interactionist perspective. The specific kind of interaction we focus on in this article is termed multilevel (or cross-level) interaction, which is a form of statistical moderation (Baron & Kenny, 1986; Raudenbush & Bryk, 2002). A classroom context acts as a moderator if relations at the individual level vary as a function of the classroom context. In relation to achievement goal theory, a classroom goal structure acts as a moderator if it either strengthens or weakens the relations between personal goals and student outcomes. A theoretical significance of the interactionist perspective is that it provides a conceptual framework for researchers to investigate whether and how the motivational system of the individual may operate differently under varying contextual conditions, thus advancing our understanding of individuals' differential vulnerability to environmental stress and differential receptivity to environmental support (Wachs, 2000). An applied significance is that findings derived from interaction analyses not only provide teachers, policy makers, and educational reformers with important information about whether an intervention works in general, but also whether it works differentially for different types of students. Accordingly, another key objective of this study was to examine how classroom goal structures moderated the relations between personal goals and student outcomes.

Of particular relevance to our interest in studying person-context interaction are studies conducted by Newman (1998) and Linnenbrink (2005). In his experimental study, Newman (1998) manipulated the goal context and randomly assigned fifth and sixth graders to either a mastery or performance goal context. He found that in the performance goal context, personal performance-approach goals tended to be negatively related to help seeking, whereas in the mastery goal context, the relation tended to be positive.

Linnenbrink (2005) conducted a quasi-experimental study in which she measured Grade 6 students' entering personal goals (mastery and performance-approach goals) and manipulated classroom goal structures by providing participating teachers with specific suggestions for practice, primarily on the basis of the evaluation and recognition components of the TARGET system (Ames, 1992; Epstein, 1988). She included three classroom goal conditions in her analysis—mastery, performance-approach, and combined mastery/performance-approach. Using multivariate analysis of variance (MANCOVA), she found no significant interactions between classroom goal conditions and person goals, and therefore concluded that "students' entering personal goal orientations did not alter the way in which they responded to the classroom goal context" (p. 204). However, Linnenbrink (2005) did not include performance-avoidance goals in her study and neither did Newman (1998). It turns out, as we show in the



Analyses and Results section, that person–context interactions were especially salient between personal performance-avoidance goals and classroom performance goal structures. To provide a comprehensive account of the patterns of person–context interactions, this study extended previous work by including personal mastery, performance-approach, and performance-avoidance goals, and by using an alternative research design and analytic strategy to study person–context interactions.

### Hypotheses

To examine whether and how classroom goal structures moderate the relations between personal goals and student outcomes, we used HLM to test three hypotheses. The first is termed the *additive hypothesis*. This is a main effect<sup>3</sup> hypothesis in which classroom goal structures and personal goals have additive contributions to the prediction of student outcomes. Under this hypothesis, for example, classroom mastery goal structures would predict outcomes at the classroom level and personal mastery goals would predict student-level outcomes, but classroom mastery goal structures do not moderate the relations between personal mastery goals and student-level outcomes.

The second hypothesis is termed the *reinforcing hypothesis*, which includes two types of interaction. In the first type of reinforcing interaction, classroom goal structures strengthen a desirable relation at the individual level (an enhancing pattern). An example consistent with this type of interaction is that classroom mastery goal structures strengthen the desirable relation between personal mastery goals and interest. Another example is that classroom performance goal structures strengthen the desirable relation between personal performance-approach goals and achievement. This type of interaction can be linked to the matching (Linnenbrink, 2004, 2005) or the person–environment fit hypothesis (Eccles et al., 1993; Hunt, 1975) proposed in prior research. The matching hypothesis suggests that a classroom goal structure that affords opportunities for the pursuit of a personal goal (i.e., a match between the person and the context) would lead to adaptive outcomes. The concept of person–environment fit or goal congruence is also central to research in organizational psychology (Edwards, Cable, Williamson, Lambert, & Shipp, 2006) and vocational choices (Holland, 1997). For example, Holland (1997) proposed that people would perceive an occupational environment to be reinforcing and satisfying when environmental features resemble their own personal characteristics. This person–environment fit would result in stability of behavior because people would receive a substantial amount of selective reinforcement of their behavior in such an environment.

Another type of reinforcing interaction is that classroom goal structures strengthen an undesirable relation at the individual level (an exacerbating pattern). An example consistent with this type of interaction is that classroom performance goal structures strengthen the undesirable relation between personal performance-avoidance goals and avoidance behavior. This pattern of interaction is consistent with the vulnerability-stress hypothesis in psychopathology (Ingram & Luxton, 2005). Vulnerability refers to a personal characteristic that predisposes an individual to a psychological disorder and stress refers to an environmental factor that disrupts the normal functioning of an individual. The hypothesis states that the probability of developing a given disorder depends

on the interaction between the degree of personal vulnerability and the level of stress experienced by the individual (Ingram & Price, 2001; Monroe & Simons, 1991). Although the vulnerability-stress hypothesis is commonly used in the field of psychopathology, it provides insight into the understanding of the reinforcing exacerbating pattern in relation to personal performance-avoidance goals. The hypothesis suggests that students who strongly endorse performance-avoidance goals would be vulnerable to a goal context that emphasizes demonstrating competence and social comparison of ability. It is worth noting that despite the similarity (or “match”) between classroom performance goal structures and personal performance-avoidance goals, this type of interaction is expected to lead to maladaptive outcomes, thus it could not be explained by the person–environment fit or matching hypothesis, which generally implies an enhancing effect.

The third hypothesis is termed the *counterbalancing hypothesis*, which also includes two types of interaction. In the first type of interaction, classroom goal structures weaken a desirable relation at the individual level (a dampening pattern). An example consistent with this type of interaction is that classroom performance goal structures weaken the desirable relation between personal mastery goals and interest. Another example is that classroom mastery goal structures weaken the desirable relation between personal performance-approach goals and achievement. In these examples, the positive potential of an individual is dampened (or not fully realized) due to goal incongruence. This pattern of interaction is consistent with the notion of “mismatch” or “misfit” in the matching (Linnenbrink, 2004, 2005) or person–environment fit hypothesis (Eccles et al., 1993; Hunt, 1975).

Another type of counterbalancing interaction is that classroom goal structures weaken an undesirable relation at the individual level (a buffering pattern). An example consistent with this type of interaction is that classroom mastery goal structures weaken the undesirable relation between personal performance-avoidance goals and avoidance behavior. This pattern of interaction is consistent with the buffering hypothesis proposed in prior research (Linnenbrink, 2004, 2005). The buffering hypothesis suggests that a supportive environment can buffer the negative impact of personal vulnerability on achievement-related outcomes.

Overall, the additive, reinforcing, and counterbalancing hypotheses proposed in this study are able to incorporate the various types of relation between classroom goal structures and personal goals. In addition to their generality and comprehensiveness, a strength of these hypotheses is that they can be tested systematically and unambiguously by specific patterns of results in HLM. An example illustrating the links between these hypotheses and specific patterns of HLM results is presented in the Analyses and Results section.

### Objectives and Research Questions

In summary, a primary objective of this study was to examine the patterns of cross-level interactions between classroom goal

<sup>3</sup> To avoid awkward and cumbersome terminology, we used the term “effects” (e.g., main effects, interactive effects, additive effects, moderating effects) in a noncausal sense to refer to predictive relations in this article.



structures (mastery and performance) and personal goals (mastery, performance-approach, and performance-avoidance) in predicting multiple outcomes in a large sample of Grade 5 students in math classrooms. Another key objective was to examine the main (or additive) effects of classroom goal structures and personal goals on predicting student outcomes, if interactions were not significant or if the patterns of interactions were ordinal (Cohen, Cohen, West, & Aiken, 2003; Pedhazur, 1997).

We addressed the following research questions in this study: (a) Did classroom goal structures moderate the student-level relations between personal goals and outcomes? (b) If cross-level interactions (or moderating effects) were significant, would they show reinforcing or counterbalancing patterns? (c) If the interactions were not significant or displayed ordinal patterns of interactions, what would be the unique contributions of classroom goal structures and personal goals to predicting outcomes? (d) Were classroom-level relations between classroom goal structures and outcomes consistent with student-level relations between personal goals and outcomes?

## Method

### Participants

This study was part of a larger research project investigating the relations between learning environments and student outcomes. The participants in this study were 3,943 Grade 5 students from 130 classrooms in 38 elementary schools in Singapore. The sample included 2,754 Chinese (69.8%), 824 Malay (20.9%), 262 Indian (6.6%), and 103 students of other ethnic origins (2.6%). The gender distribution of the sample was 52.7% male and 47.3% female. The mean age of the participants was 11.4 years. English is the medium of instruction in Singapore, and all students formally start learning English in Grade 1.

### Sampling Design and Procedure

Low statistical power to detect interaction effects, especially in the presence of measurement errors, has been an issue of great concern (Aiken & West, 1991; Cohen et al., 2003; McClelland & Judd, 1993). Therefore, consideration of sampling design is especially important in testing interaction hypotheses. In this study, we sampled a large number of students and classrooms using a stratified random sampling technique. Schools were divided into three strata based on their prior aggregate school achievement and 13 schools were randomly selected from each stratum. One school dropped out of the study, leaving a total of 38 schools in the sample. About half of the Grade 5 classrooms in each participating school were randomly selected to do the math survey and assessment. Our sampling design ensured that we tapped sufficient "natural variance" of classroom characteristics and student demography by selecting schools and classrooms that covered a broad spectrum of achievement levels.

The procedure consisted of two parts. Part 1 was an online survey conducted in the 8th month from the beginning of the school year. The online survey included two forms. Half of the students in each class were randomly selected to complete survey Form 1, in which students reported their motivational beliefs and achievement-related behaviors in their math classrooms. The other

half of the students in the same class completed survey Form 2, in which students reported their perceived classroom goal structures. In other words, half of the students provided student-level (or level-1) data and the other half provided classroom-level (or level-2) data. In effect, half of the students served as independent raters of classroom goal structures. Because different groups of students provided data at different levels, potential inflation of cross-level interactions would be reduced. Such inflation could occur, for example, as a result of similar item wordings of the personal goal and classroom goal structure measures. The average numbers of students completing Forms 1 and 2 per class were 14.8 and 15.5, respectively. In Part 2 of the study, which was conducted about 1 month after the survey administration, a math achievement test was administered to all the students who had completed either Form 1 or Form 2 of the survey.

### Measures

All of the items on the survey were rated on 5-point Likert scales (1 = *strongly disagree* to 5 = *strongly agree* or 1 = *never* to 5 = *always*). Sample items of self-report scales are provided in the Appendix.

**Classroom goal structures.** Two types of classroom goal structures were assessed—classroom mastery and classroom performance goal structures. The classroom mastery goal structure describes an environment in which the teacher emphasizes that learning, task mastery, and working hard are important. The classroom performance goal structure describes an environment in which the teacher emphasizes that demonstrating high ability and getting better grades than other students are important. It is worth noting that our measure of the classroom performance goal structure focuses on the approach dimension. The measures of classroom mastery and classroom performance goal structures were adapted from the Patterns of Adaptive Learning Survey (PALS; Midgley et al., 2000).

A confirmatory factor analysis was conducted to examine the factor structure of the constructs. A two-factor structure provided a good fit for the data,  $\chi^2(26, N = 2017) = 239.61$ , Tucker-Lewis index (TLI) = .92, comparative fit index (CFI) = .95, root-mean-square error of approximation (RMSEA) = .06. Each scale showed adequate internal consistency ( $\alpha = .83$  for classroom mastery goal structure and  $\alpha = .73$  for classroom performance goal structure).

Classroom-level measures of classroom goal structures were derived from aggregating (i.e., averaging within each classroom) individual students' perceptions of classroom goal structures (Karabenick, 2004; A. M. Ryan, Gheen, & Midgley, 1998). These aggregated measures were used as level-2 predictors in HLM. The total variability of perceived classroom goal structures consisted of the within-class and between-class components. The aggregated measures we used reflect the between-class component.

**Personal goals.** Three types of personal goals were measured: personal mastery, personal performance-approach, and personal performance-avoidance goals. The personal mastery goal scale assessed students' desire to learn new things and to master challenging concepts in math. The personal performance-approach goal scale assessed students' desire to demonstrate their superior ability relative to their peers and to obtain favorable judgment from teachers. The personal performance-avoidance goal scale



assessed students' desire to hide their weaknesses in math and to avoid being perceived as incompetent by their teachers and peers. The measures of personal goals were adapted from PALS (Midgley et al., 2000).

A confirmatory factor analysis was conducted to examine the factor structure of the constructs. A three-factor structure provided an adequate fit for the data,  $\chi^2(62, N = 1926) = 707.91$ ,  $TLI = .91$ ,  $CFI = .94$ ,  $RMSEA = .07$ . Each scale showed adequate to high internal consistency ( $\alpha = .85$  for personal mastery goal,  $\alpha = .87$  for personal performance-approach goal, and  $\alpha = .72$  for personal performance-avoidance goal).

**Achievement.** A multiple-choice math achievement test was developed for this study because a standardized test of math achievement at Grade 5 was not available in Singapore. The test was intended to assess students' knowledge and skills in math at the Grade 5 level. It included four types of questions, which required understanding basic math concepts, performing routine procedures, using complex procedures, and solving novel problems. A panel of researchers and school teachers who had experience teaching math constructed and reviewed the items to ensure the content validity, clarity, and grade-level appropriateness of the assessment instrument in the local context. A pilot study was conducted to select items from the item pool on the basis of their psychometric quality such as item difficulty, item discrimination, and functioning of distractors. A final set of 27 items was selected and administered in this (main) study. To select items for final scoring, we adopted the criterion used in the Trends in Mathematics and Science Study 1999 (TIMSS 1999; Mullis & Martin, 2000). Four items that had an item discrimination index less than 0.2 were dropped. The 23 items that were used for final scoring had high reliability ( $\alpha = .87$ ). Standardized IRT (item response theory) scores were used in further analyses.

**Adaptive and maladaptive motivational outcomes.** Two adaptive and two maladaptive motivational outcomes were assessed in this study. The adaptive motivational outcomes included students' engagement and interest in math classes. Our measure of engagement was based on students' report of their attention, effort, and participation in their math classes (Steinberg, Lamborn, Dornbusch, & Darling, 1992; Wellborn & Connell, 1987). Our measure of interest was based on students' reports of their intrinsic motivation and enjoyment in their math classes (Elliot & Church, 1997). The maladaptive motivational outcomes included avoidance coping and effort withdrawal. The avoidance coping scale assessed students' tendency to give up when the work was difficult or boring. It was adapted from the Motivated Strategies for Learning Questionnaire (Pintrich, Smith, Garcia, & McKeachie, 1993). The effort withdrawal scale assessed students' tendency to hold back or minimize effort in their math work (Meece, Blumenfeld, & Hoyle, 1988; Nicholls, Patashnick, & Nolen, 1985).

A confirmatory factor analysis was conducted to examine the factor structure of the constructs. A second-order factor model, in which interest and engagement loaded onto the higher order factor of adaptive functioning and effort withdrawal and avoidance coping loaded onto the higher order factor of maladaptive functioning, provided a good fit for the data,  $\chi^2(99, N = 1926) = 552.89$ ,  $TLI = .96$ ,  $CFI = .97$ ,  $RMSEA = .05$ . Each scale showed adequate to high internal consistency ( $\alpha = .93$  for interest,  $\alpha = .74$  for engagement,  $\alpha = .80$  for effort withdrawal, and  $\alpha = .79$  for avoidance coping).

## Analyses and Results

### *Analytic Approach to Modeling Student Outcomes*

All predictors and outcome variables were standardized before running hierarchical linear modeling (HLM). Outcome variables and level-1 predictors were standardized at level 1. Standardized level-2 predictors were derived from first aggregating level-1 scores to level 2 and then standardizing the level-2 scores at level 2. One-way analysis of variance (ANOVA) with random effects model (the unconditional model or Model 0) was used to estimate the proportion of within- and between-class variances in the outcome variables (Raudenbush & Bryk, 2002).

The intraclass correlation coefficient (ICC) measures the proportion of total variance in an outcome variable explained by between-class differences. ICC was 62.0% for math achievement, 7.6% for engagement, 4.1% for interest, 9.4% for effort withdrawal, and 5.8% for avoidance coping.<sup>4</sup> Chi-square tests were performed to examine the significance of between-class variances in the unconditional models. We found that between-class variances were significant for all of the five outcome variables,  $\chi^2$ s (129,  $N = 1926$ ) = 219.61 to 3016.18,  $ps < .001$ .

The next HLM analysis was performed to evaluate the predictive relations of personal goals to student outcomes. A student-level or level-1 model (Model 1) was run to examine the statistical significance of the three level-1 predictors: personal mastery, performance-approach, and performance-avoidance goals. To determine whether the three slopes were fixed or random, we performed a multivariate likelihood-ratio test for each of the outcome variables. The deviance statistic associated with the full random-coefficient regression model, in which all the three slopes were random, was compared with the corresponding statistic associated with the restricted model, in which all the three slopes were fixed. The purpose of the multivariate procedure is to test the omnibus null hypothesis and to protect against the inflation of Type I error rates. If the multivariate likelihood-ratio test was not significant, all the three slopes were fixed. If it was significant, univariate  $\chi^2$  tests would then be performed. The univariate test served as a post hoc procedure to identify which specific slopes were random. If univariate  $\chi^2$  tests were not significant, the slopes of the corresponding level-1 predictors were fixed (Raudenbush & Bryk, 2002). On the basis of the sequential procedures, slopes were either identified to be fixed or random. A fixed slope indicates that level-1 relations are homogeneous across classrooms, whereas a random slope indicates that level-1 relations vary across classrooms. The variation of slopes across classrooms represents a source of variance to be explained by classroom-level variables. Therefore, only random slopes were modeled in further analyses by adding level-2 predictors (see Table 2). In this study, the slopes relating personal mastery goals to avoidance coping, personal performance-approach goals to interest, personal performance-avoidance goals to engagement, personal performance-avoidance

<sup>4</sup> ICC was 8.3% for classroom mastery goal structures and 12.3% for classroom performance goal structures. ICC's for the predictors are reported for descriptive purposes only, because it is not our aim to explain the variance components associated with the predictors.

goals to effort withdrawal, and personal performance-avoidance goals to avoidance coping were found to be random.

Finally, classroom mastery and classroom performance goal structures (the level-2 predictors) were added to build the full model (Model 2). Model 2 was used to test cross-level interactions between classroom goal structures and personal goals as well as their unique main effects:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(\text{PMG}) + \beta_{2j}(\text{PPAP}) + \beta_{3j}(\text{PPAV}) + r_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(\text{CMG}) + \gamma_{02}(\text{CPG}) + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(\text{CMG}) + \gamma_{12}(\text{CPG}) + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}(\text{CMG}) + \gamma_{22}(\text{CPG}) + u_{2j}$$

$$\beta_{3j} = \gamma_{30} + \gamma_{31}(\text{CMG}) + \gamma_{32}(\text{CPG}) + u_{3j}$$

$Y_{ij}$  is the outcome variable, PMG is the personal mastery goal, PPAP is the personal performance-approach goal, PPAV is the personal performance-avoidance goal, CMG is the classroom mastery goal structure, and CPG is the classroom performance goal structure. The equation above represents the most elaborate form of Model 2. If a slope was fixed, the coefficients associated with CMG and CPG would be set to zero (i.e., no level-2 predictors would be added).

The cross-level interaction between classroom goal structures and personal goals can be interpreted as a statistical moderation effect (Baron & Kenny, 1986; Raudenbush & Bryk, 2002). The reinforcing hypothesis is supported if the interaction term is significant and its sign is the same as that of the main effect term for personal goals, the counterbalancing hypothesis is supported if the interaction term is significant and its sign is opposite to that of the main effect term for personal goals, and the additive hypothesis is supported if the interaction term is not significant. As an illustration, in Model 2,  $\gamma_{10}$  is the coefficient for the main effect of personal mastery goals on the outcome variable,  $\gamma_{11}$  is the coefficient for Classroom Mastery Goal Structure  $\times$  Personal Mastery Goal interaction, and  $\gamma_{12}$  is the coefficient for Classroom Performance Goal Structure  $\times$  Personal Mastery Goal interaction. In addition,  $\gamma_{01}$  and  $\gamma_{02}$  can be interpreted as the main effects of classroom mastery and classroom performance goal structures, respectively. A reinforcing effect of classroom mastery goal structures on the relation be-

tween personal mastery goals and the outcome variable is demonstrated if, for example,  $\gamma_{11}$  and  $\gamma_{10}$  are of the same sign, whereas a counterbalancing effect of classroom mastery goal structures is demonstrated if  $\gamma_{11}$  and  $\gamma_{10}$  are of opposite signs. Classroom mastery goal structures do not interact with personal mastery goals if  $\gamma_{11}$  is not significantly different from zero.

We also estimated the percentage of variance explained as a result of adding predictors in successive models. Besides conceptual considerations in relation to our research objectives, the sequence of model building is based on Raudenbush and Bryk's (2002) recommendation on the proper use of variance-explained statistics—"the variance explained in a level-2 parameter, such as  $\beta_{0j}$ , is conditional on a fixed level-1 specification" (p. 150). Thus, the comparison model (or reference model) we used to compute the variance explained statistic is always a nested model of the more complex model that follows (i.e., Model 2 vs. Model 1 or Model 1 vs. Model 0).

### Descriptive Statistics and Zero-Order Correlations

Descriptive statistics and zero-order correlations among the student-level variables used in this study are presented in Table 1. The zero-order correlation between the two level-2 predictors, classroom mastery goal structures and classroom performance goal structures, was not significant ( $r = -.053, p > .05, n = 130$ ).

### Cross-Level Interactions Between Classroom Goal Structures and Personal Goals

Table 2 presents the parameter estimates of fixed effects and results of hypothesis tests for the full model (Model 2). The most notable cross-level interactions were found between classroom performance goal structures and personal performance-avoidance goals. Classroom performance goal structures had significant moderating effects on the relation between personal performance-avoidance goals and engagement ( $\gamma = -.049, p < .05$ ), between performance-avoidance goals and effort withdrawal ( $\gamma = .101, p < .01$ ), and between performance-avoidance goals and avoidance coping ( $\gamma = .090, p < .01$ ).

The moderating effects of classroom performance goal structures on level-1 relations are shown in Figures 1–3. In classrooms with stronger emphasis on performance goals, personal performance-avoidance goals tended to be more negatively related to engagement,

Table 1  
Descriptive Statistics and Zero-Order Correlations Among Student-Level Variables

	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	7	8
Student level ( $n = 1,926$ )										
1. Achievement	0.00	1.00	—							
2. Engagement	4.16	.66	.19	—						
3. Interest	4.00	.94	.19	.32	—					
4. Effort withdrawal	2.07	.97	-.29	-.25	-.22	—				
5. Avoidance coping	1.98	.98	-.24	-.24	-.31	.48	—			
6. Personal mastery goal	3.99	.75	.19	.43	.77	-.24	-.30	—		
7. Personal performance-approach goal	3.47	1.03	-.01 <sup>†</sup>	.17	.25	.07	-.02 <sup>†</sup>	.31	—	
8. Personal performance-avoidance goal	2.78	1.02	-.27	-.09	-.01 <sup>†</sup>	.41	.25	.01 <sup>†</sup>	.38	—

Note. All correlations are significant at the .01 level, except those marked with a dagger †, which are not significant at .05 level.



Table 2  
Results From HLM Analyses Predicting Achievement and Motivational Outcomes

Variable	Achievement			Engagement			Interest			Effort withdrawal			Avoidance coping			
	Fixed effect	$\gamma$	SE	p	$\gamma$	SE	p	$\gamma$	SE	p	$\gamma$	SE	p	$\gamma$	SE	p
$\beta_0$ Intercept																
$\gamma_{00}$ Mean		-.126	.060	.038	-.008	.026	.757	.000	.017	.988	-.003	.023	.893	.005	.022	.814
$\gamma_{01}$ Classroom mastery		.129	.060	.034	.052	.029	.078	.024	.018	.182	-.066	.027	.017	-.097	.024	<.001
$\gamma_{02}$ Classroom performance		-.385	.055	<.001	-.077	.027	.005	.014	.019	.462	.126	.025	<.001	.099	.026	<.001
$\beta_1$ Personal mastery slope																
$\gamma_{10}$ Mean		.140	.016	<.001	.401	.026	<.001	.764	.018	<.001	-.259	.022	<.001	-.300	.024	<.001
$\gamma_{11}$ Classroom mastery		— <sup>a</sup>	—	—	—	—	—	—	—	—	—	—	—	-.038	.025	.133
$\gamma_{12}$ Classroom performance		—	—	—	—	—	—	—	—	—	—	—	—	-.022	.026	.400
$\beta_2$ Personal performance-approach slope																
$\gamma_{20}$ Mean		.001	.019	.971	.085	.023	<.001	.016	.019	.395	-.012	.023	.606	-.014	.024	.556
$\gamma_{21}$ Classroom mastery		—	—	—	—	—	—	-.006	.016	.713	—	—	—	—	—	—
$\gamma_{22}$ Classroom performance		—	—	—	—	—	—	.013	.016	.426	—	—	—	—	—	—
$\beta_3$ Personal performance-avoidance slope																
$\gamma_{30}$ Mean		-.094	.018	<.001	-.102	.023	<.001	-.021	.016	.204	.386	.029	<.001	.237	.026	<.001
$\gamma_{31}$ Classroom mastery		—	—	—	-.033	.024	.179	—	—	—	-.003	.025	.911	.000	.025	.990
$\gamma_{32}$ Classroom performance		—	—	—	-.049	.024	.042	—	—	—	.101	.028	.001	.090	.024	<.001

Note. Coefficients significant at the .05 level are boldfaced. Number of level-1 units = 1,926; number of level-2 units = 130.  
<sup>a</sup> Classroom mastery and classroom performance goal structures were not entered as level-2 predictors if the slopes of the corresponding personal goals were fixed.

more positively related to effort withdrawal, and more positively related to avoidance coping. The patterns can be characterized as ordinal interactions, in which nonparallel lines do not cross over within the range of interest (Cohen et al., 2003; Pedhazur, 1997).  
Comparing Model 2 (the full model) with Model 1 (the student-level model), we found that classroom goal structures accounted for 20.2% of the between-class variance in the slope of performance-

avoidance goals predicting engagement, 16.8% in the slope of performance-avoidance goals predicting effort withdrawal, and 25.91% in the slope of performance-avoidance goals predicting avoidance coping.  
No interactions were found between classroom goal structures (either mastery or performance) and personal mastery goals and between classroom goal structures and personal performance-

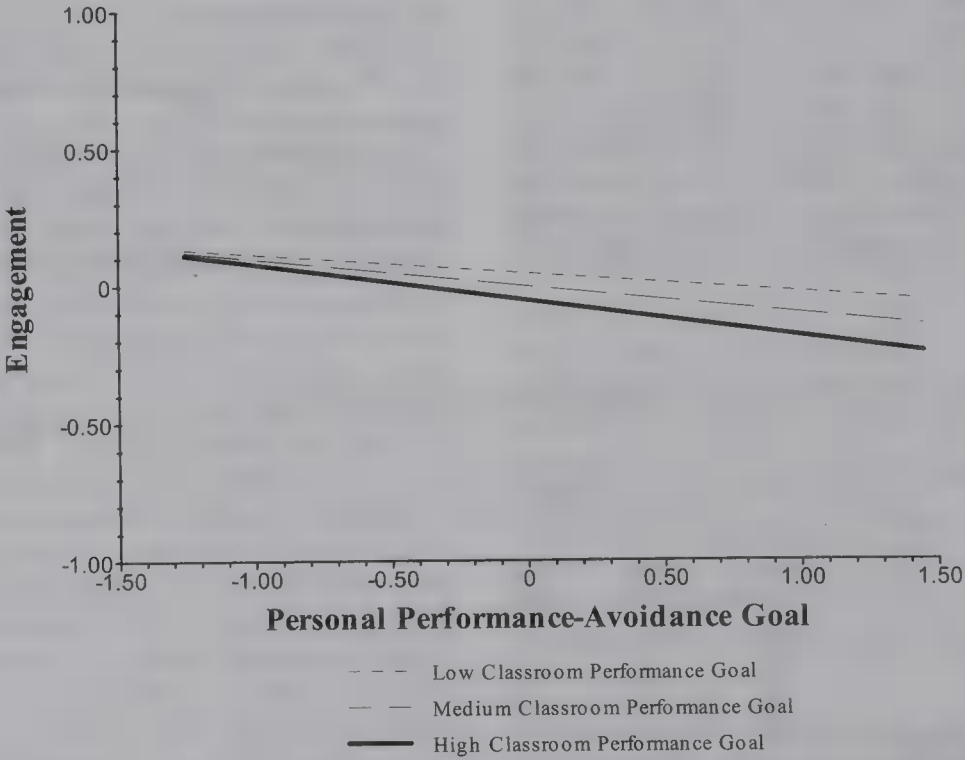


Figure 1. Interaction between classroom performance goal structure and personal performance-avoidance goal predicting engagement.

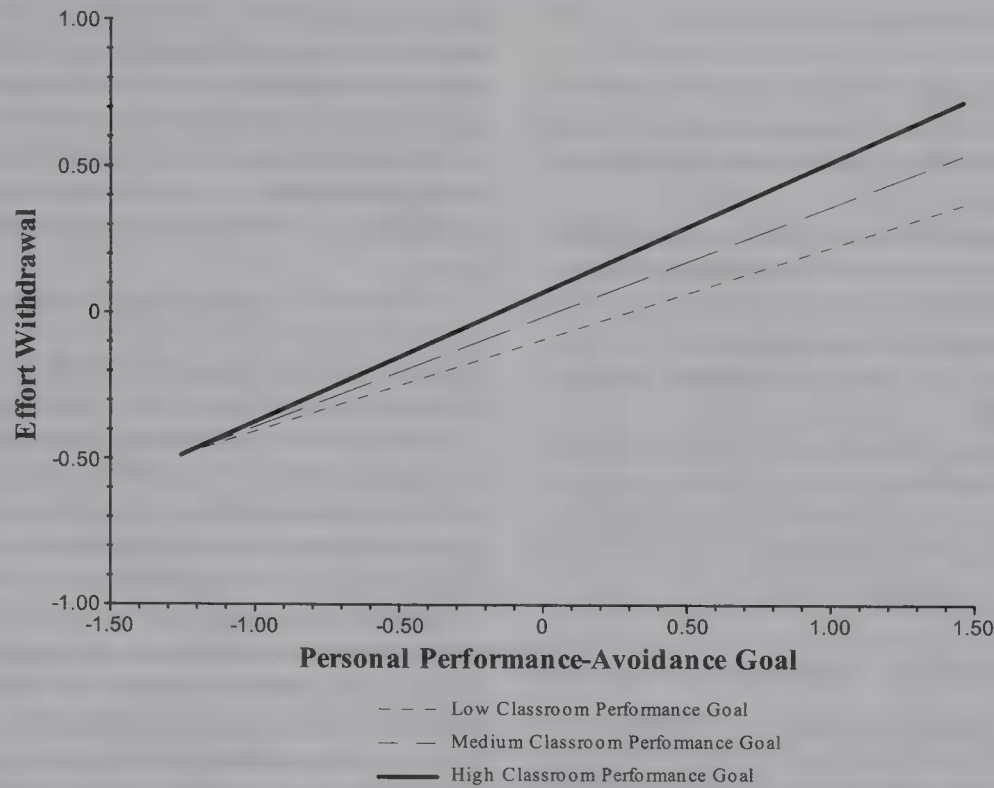


Figure 2. Interaction between classroom performance goal structure and personal performance-avoidance goal predicting effort withdrawal.

approach goals. For variables that produced interactions, the interaction pattern was of the ordinal type, suggesting that the main effects can be used to summarize the overall predictor-outcome relations. Therefore, we proceed to present results on the main effects of personal goals and classroom goal structures whether or not interactions were found.

#### Main Effects of Personal Goals

As shown in Table 2, personal mastery goals were positively related to math achievement ( $\gamma = .140, p < .001$ ), engagement ( $\gamma = .401, p < .001$ ), and interest ( $\gamma = .764, p < .001$ ). Moreover, personal mastery goals were negatively related to effort with-

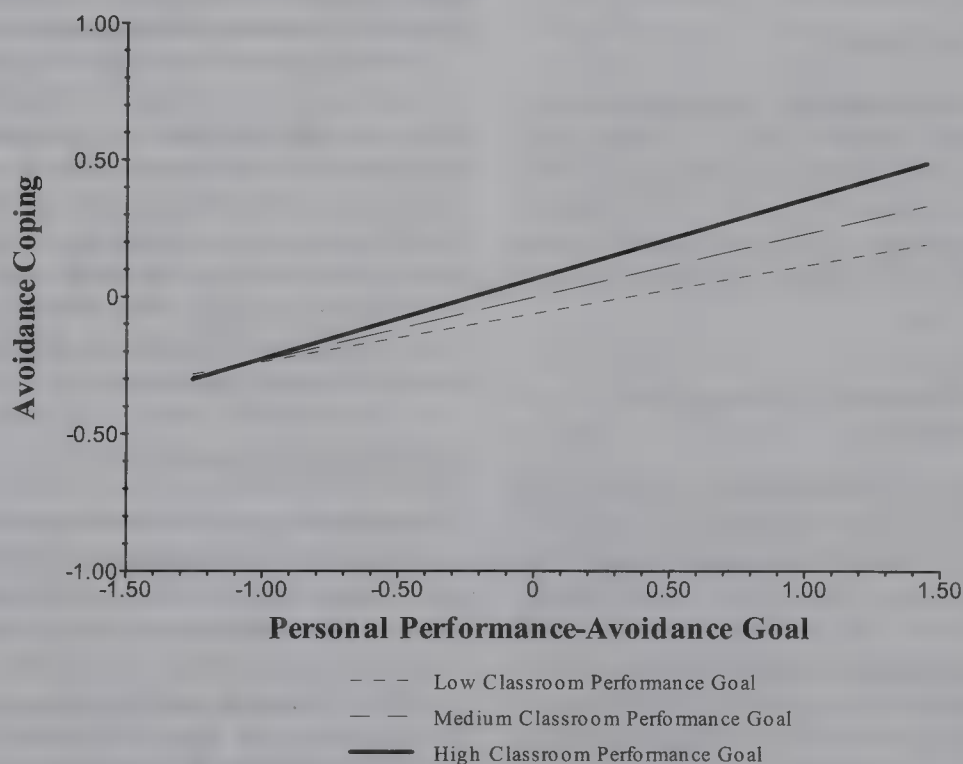


Figure 3. Interaction between classroom performance goal structure and personal performance-avoidance goal predicting avoidance coping.



drawal ( $\gamma = -.259, p < .001$ ) and avoidance coping ( $\gamma = -.300, p < .001$ ).

Personal performance-approach goals were positively related to engagement ( $\gamma = .085, p < .001$ ). For other outcomes, personal performance-approach goals were not significant predictors at the .05 level.

Personal performance-avoidance goals were negatively related to math achievement ( $\gamma = -.094, p < .001$ ) and engagement ( $\gamma = -.102, p < .001$ ), but unrelated to interest ( $\gamma = -.021, p > .05$ ). Moreover, personal performance-avoidance goals were positively related to effort withdrawal ( $\gamma = .386, p < .001$ ) and avoidance coping ( $\gamma = .237, p < .001$ ).

Comparing Model 1 (the student-level model) with Model 0 (the unconditional model), we found that personal goals accounted for 5.9% of the within-class (or level-1) variance for math achievement, 19.1% for engagement, 59.4% for interest, 25.7% for effort withdrawal, and 16.6% for avoidance coping.

### *Main Effects of Classroom Goal Structures*

As shown in Table 2, classroom mastery goal structures were positively related to math achievement ( $\gamma = .129, p < .05$ ), but negatively related to effort withdrawal ( $\gamma = -.066, p < .05$ ) and avoidance coping ( $\gamma = -.097, p < .001$ ). In contrast, classroom performance goal structures were negatively related to math achievement ( $\gamma = -.385, p < .01$ ) and engagement ( $\gamma = -.077, p < .01$ ) but positively related to effort withdrawal ( $\gamma = .126, p < .001$ ) and avoidance coping ( $\gamma = .099, p < .001$ ).

Comparing Model 2 (the full model) with Model 1 (the student-level model), we found that classroom goal structures accounted for 27.2% of the between-class (or level-2) variance for math achievement, 14.6% for engagement, 10.5% for interest, 40.3% for effort withdrawal, and 56.8% for avoidance coping, after controlling for personal goals at level 1.

## Discussion

The main purposes of this study were to examine how classroom goal structures moderated the predictive relations between personal goals and student outcomes, as well as how these variables operated additively at the individual and classroom levels. Specifically, we used HLM to test three hypotheses: the reinforcing hypothesis, the counterbalancing hypothesis, and the additive hypothesis.

### *Cross-Level Interactions Between Classroom Goal Structures and Personal Goals*

The reinforcing hypothesis has received empirical support from the findings of cross-level interactions between classroom performance goal structures and personal performance-avoidance goals in predicting engagement, effort withdrawal, and avoidance coping. We found that a strong focus on classroom performance goal structures tended to reinforce (or exacerbate) the negative association between personal performance-avoidance goals and student engagement. The reinforcing effects are also evident in predicting maladaptive motivational outcomes. High levels of classroom performance goal structures tended to reinforce (or exacerbate) the positive predictive relations of personal performance-avoidance

goals to both effort withdrawal and avoidance coping. In other words, our finding suggests that in classrooms that emphasize demonstrating ability, social comparison of performance, and competition for good grades, students who are oriented toward performance-avoidance goals would be less likely to be engaged in their math classes, more likely to withdraw their effort in learning math, and more likely to give up when the work is difficult or boring.

The additive hypothesis was supported by the finding that classroom performance goal structures did not moderate the predictive relations of either personal mastery goals or performance-approach goals to any of the outcome variables. It was also supported by the finding that classroom mastery goal structures did not interact with any of the three personal goals in predicting any of the outcome variables, which indicated that classroom mastery goal structures operated in an additive way with personal goals to predict achievement and motivational outcomes.

Our findings are by and large consistent with Linnenbrink's (2005), despite differences between the two studies in numerous methodological and sample characteristics. In her quasi-experimental study, she found that classroom goal conditions did not interact with either personal mastery goals or personal performance-approach goals. Linnenbrink, however, did not include personal-avoidance goals in her study, which turn out to be critical for testing the interaction hypotheses in this study.

This study extends previous research in two ways. First, our findings triangulate Linnenbrink's (2005) findings with respect to personal mastery and personal performance-approach goals by using an alternative research design and analytic approach. This approach allows us to associate specific HLM results with specific hypotheses in a systematic way. Second, including personal performance-avoidance goals in testing the interaction hypotheses provides a richer and more complete picture of person-context interactions by demonstrating the moderating effects of classroom performance goal structures on the maladaptive relations between personal performance-avoidance goals and motivational outcomes.

Overall, we obtained corroborative evidence for the additive and reinforcing hypotheses, but no evidence for the counterbalancing hypothesis. Whether and how classroom goal structures interact with personal goals depend on the specific combination of classroom goal structures and personal goals and the specific outcome variables used in the analysis. Before we elaborate on the implications of our findings for educational practices, we first discuss and interpret the main effect findings because it is important to take into account multiple levels of analysis in making recommendations for educational practices.

### *Main Effects of Personal Goals*

We found that personal mastery goals were positively related to math achievement, engagement, and interest in math, but negatively related to effort withdrawal and avoidance coping. The patterns for performance-avoidance goals were quite consistent, but in the opposite direction—performance-avoidance goals were negatively related to math achievement and engagement, but positively related to effort withdrawal and avoidance coping. Performance-approach goals were positively related to engagement, but unrelated to the other outcome variables. Overall, the main effects of personal goals are consistent with the following



general conclusions derived from prior research: (a) personal mastery goals are adaptive, (b) performance-avoidance goals are maladaptive; and (c) evidence for the relations between performance-approach goals and outcomes is mixed and inconclusive (Elliot, 2005; Harackiewicz et al., 2002; Midgley et al., 2001). The demonstration of consistent results based on an Asian sample enhances the cross-cultural generalizability of the achievement goal theory.

### *Main Effects of Classroom Goal Structures*

Classroom mastery goal structures were positive predictors of math achievement and negative predictors of effort withdrawal and avoidance coping. In other words, in classrooms where teachers were perceived to emphasize learning and improving, students tended to perform better in math tests, were less likely to withdraw effort in their math work, and were less likely to give up when the work was difficult or boring. In contrast, classroom performance goal structures were negative predictors of math achievement and engagement, but positive predictors of effort withdrawal and avoidance coping. That is, in classrooms with stronger emphasis on demonstrating one's competence, social comparison of ability, and competition for high grades, students tended to perform worse in math tests, were less likely to pay attention and participate in group work in class, were more likely to withdraw effort in their math work, and were more likely to give up when the work was difficult or boring.

Prior research that adopted a multilevel framework and used HLM procedures has documented similar patterns of results,<sup>5</sup> especially for maladaptive outcomes. For example, classroom mastery goal structures were found to be significant negative predictors of self-handicapping (Turner et al., 2002), avoidance of help-seeking (Karabenick, 2004; A. M. Ryan et al., 1998; Turner et al., 2002), and disruptive behaviors (Kaplan, Gheen, et al., 2002). In contrast, classroom performance goal structures were positively related to self-handicapping (Urdan, Midgley, & Anderman, 1998), avoidance of help-seeking (Karabenick, 2004; A. M. Ryan et al., 1998), and disruptive behavior (Kaplan, Gheen, et al., 2002). This study provides corroborative evidence for the role of classroom mastery goal structures and classroom performance goal structures in students' achievement and motivational outcomes in an Asian context. In addition, by including both adaptive and maladaptive outcomes in the same study, we provide a comprehensive set of findings to document differential relations for different types of classroom goal structure.

Comparisons of level-1 and level-2 relations reveal that personal mastery goals and classroom mastery goal structures had similar patterns of predictive relations to outcomes. Surprisingly, classroom performance goal structures, despite the focus on the approach dimension, were consistent with personal performance-avoidance goals (but not with personal performance-approach goals) in predicting outcomes at their respective levels of analysis. These findings have important implications for practice, which we elaborate in the next section.

### *Implications for Classroom Practices*

Our study underscores the importance of studying achievement motivation in a multilevel framework. This point is particularly pertinent in light of recent debates about whether personal perfor-

mance-approach goals are adaptive or not (Harackiewicz et al., 2002; Kaplan & Middleton, 2002; Midgley et al., 2001). In the current study, personal performance-approach goals were positively related to engagement and were unrelated to other outcome variables. From the individual level of analysis, it seems harmless or even beneficial to emphasize the approach dimension of performance goals. However, this finding alone cannot be directly translated into recommendations for educational practices because interventions that involve changing teachers' classroom practices are usually conducted at the classroom level. Classroom goal structures could have additive effects on student outcomes independent of their personal goals, as well as moderating effects on the relations between personal goals and student outcomes. We therefore caution against making recommendations for classroom practices on the basis of research conducted solely at the individual level of analysis (Roeser, 2004).

In line with the above argument, we take into account multilevel relations in drawing our conclusions. Both the interaction and main effect findings regarding the maladaptive role of classroom performance goal structures suggest that teachers should deemphasize performance goals in their classrooms, even if teachers' goal messages are meant to encourage students to demonstrate superior ability and to get higher grades than their peers. As our findings indicate, classroom performance goal structures are negatively related to math achievement and engagement, and positively related to effort withdrawal and avoidance coping. What is worrisome is that classroom performance goal structures appear to be especially detrimental to students who strongly endorse performance-avoidance goals, due to the reinforcing or exacerbating effects on the maladaptive relations between performance-avoidance goals and outcomes. Given the robust and consistent empirical generalization of the negative impact of personal performance-avoidance goals on student outcomes (see Elliot, 2005; Pintrich & Schunk, 2002; Urdan, Ryan, Anderman, & Gheen, 2002, for reviews), students who are oriented toward performance-avoidance goals are especially at-risk for maladaptive outcomes. Emphasizing performance goals in the classroom may further exacerbate the risk to this group of students, even if teachers focus on the approach dimension of classroom performance goal structures.

In addition, given the convergence of evidence from the current study and prior research (Kaplan, Gheen, et al., 2002; Karabenick, 2004; A. M. Ryan et al., 1998; Turner et al., 2002) regarding the adaptive role of both classroom mastery goal structures and personal mastery goals, it is advisable for teachers to place greater emphasis on mastery goals in their classrooms. Our findings show that classroom mastery goal structures and personal mastery goals operate in an additive way and the patterns of predictor-outcome relations are similar at both the classroom-level and student-level of analysis.

<sup>5</sup> A number of prior studies investigated the relations between classroom goal structures and outcomes at a single level (student-level) of analysis (e.g., Anderman & Midgley, 1997; Church et al., 2001; Kaplan & Maehr, 1999; Roeser et al., 1996). Because the within-class and between-class components of students' perceptions of classroom goal structures were not separated in these studies, it is difficult to compare their results with ours.



## Limitations

Several limitations of this study should be noted. First, the use of correlational data does not permit us to infer causal flows from classroom goal structures and personal goals to student outcomes. Nevertheless, the evidence that supports the additive hypothesis (i.e., classroom mastery goal structures and classroom performance goal structures did not interact with either personal mastery goals or personal performance-approach goals) shows remarkable consistency with prior research using a quasi-experimental design (Linnenbrink, 2005). We believe that our findings and Linnenbrink's are complementary and mutually corroborative. Moreover, we extend previous research by demonstrating cross-level interactions between classroom performance goal structures and personal performance-avoidance goals. This finding, however, needs to be triangulated by quasi-experimental or experimental studies. It is hoped that our study will stimulate further research that uses alternative research designs to fully understand the complex interplay between personal goals and classroom goal structures.

Second, our measure of classroom performance goal structures primarily assessed the approach dimension. But it is important to note that, surprisingly, our "approach measure" had significant interactions with personal performance-avoidance goals, but not with personal performance-approach goals. Moreover, classroom performance goal structures were consistent with personal performance-avoidance goals (but not with personal performance-approach goals) in predicting outcomes at their respective levels of analysis. Thus, the fact that we focused on the approach dimension of classroom performance goal structures actually strengthens the case against emphasizing performance goals in the classroom. Students who strongly endorse performance-avoidance goals may not necessarily interpret the "approach" aspect of the performance goal messages despite the intentions of their teachers. It remains an empirical question whether students can distinguish between the approach and avoidance dimensions of classroom goal structures and, if they do, how they would interpret the goal messages conveyed to them. Karabenick (2004) reported a zero-order correlation of .91 between aggregated student perceptions of classroom performance-approach goals and classroom performance-avoidance goals, whereas Kaplan, Gheen, et al. (2002) reported the corresponding correlation to be .41. A comprehensive and systematic validity study would be needed to determine the empirical viability and utility of distinguishing between the approach and avoidance dimensions of classroom performance goal structures.

Third, our measures of classroom goal structures were based on students' self-reports, which were their subjective interpretations of goal messages conveyed by their teachers. There are disagreements in the literature about the merits of using student perceptions and interpretations of goal messages (or the functional significance of the goal context) are more important than the objective reality of classroom contexts in influencing students' achievement-related behaviors (Ames, 1992; Meece et al., 2006; A. M. Ryan et al., 1998; R. M. Ryan & Grolnick, 1986). When both teachers' reports and students' reports of classroom goal structures were used in the same study, aggregated student-report measures tended to be more strongly related to outcomes than did teacher-report measures (Kaplan, Gheen, et al., 2002; A. M. Ryan et al., 1998; Urdan et al., 1998). Moreover, there is evidence

showing that students' self-reports of classroom goal structures were systematically related to teachers' practices based on classroom observations (Patrick, Anderman, Ryan, Edelin, & Midgley, 2001). Furthermore, from a measurement perspective, aggregation of student perceptions reduces measurement and other unsystematic errors and hence produces more reliable measures than teachers' reports or classroom observations based on a single observer.

However, a limitation of student perceptions is that it is difficult to make recommendations for classroom practices without understanding how objective classroom environments impact student outcomes (Linnenbrink, 2005; Urdan, 1997, 2004). Experimental and quasi-experimental studies of how manipulations of objective goal-related features in the classroom may alter students' personal goals and achievement-related outcomes would provide practitioners with concrete guidelines and specific practices for altering classroom goal structures. In addition, there is a need for more research that uses multiple methods such as classroom observations, interviews, simulated recalls, and experimental manipulations to triangulate and validate findings based on student perceptions (Urdan, 2004).

A fourth issue in relation to achievement goal research is the stability of personal goals. Experimental manipulations have been found to be effective in eliciting or changing particular personal goals in the laboratory setting, at least briefly, suggesting that personal goals could be responsive to situational demands and cues (Barron & Harackiewicz, 2001; Elliot & Harackiewicz, 1996; Elliott & Dweck, 1988). But evidence based on longitudinal survey research has shown that personal goals are relatively stable over time (Anderman & Midgley, 1997; Midgley et al., 1998; Wolters, Yu, & Pintrich, 1996). The mixed evidence base has led some authors to conclude that personal goals could display both dispositional and situational characteristics, depending on the research context (Elliot, 2005; Urdan, 1997, 2004). In experimental settings, goal messages are usually unambiguous and situational demands are strong, whereas in natural classroom environments, goal messages are usually mixed and the situational demands are weaker than those in experimental settings (Urdan, 1997). Although we expected personal goals to be relatively stable as this study was conducted in natural classroom settings, it is possible that classroom goal structures would have had time to influence personal goals as the data were collected in the 8th month from the beginning of the school year.

A limitation of this study is that a cross-sectional design does not allow us to empirically assess the stability of personal goals over time. This design could only capture the product of a series of potentially complex processes at a particular point in time, but the specific processes involved remain unclear. For example, although the current study demonstrates that, at the time of data collection, the relation between personal performance-avoidance goals and avoidance coping tended to be stronger in classrooms with higher levels of performance goal structures (a reinforcing-exacerbating pattern), the specific processes leading to this pattern of person-context interaction remain unclear. If classroom goal structures influenced personal goals, which in turn influenced achievement-related outcomes, the product of such mediating processes, together with those of any other processes, would be registered as the final product (statistical relations) at the point of data collection. Evidence on the stability (or instability) of personal goals would allow researchers to assess the plausibility of the mediation expla-



nation.<sup>6</sup> Future research that employs a longitudinal design to track students' personal goals over time would help to shed light on the processes involved (Anderman et al., 1999; Anderman & Midgley, 1997).

<sup>6</sup> It is important to note that mediation and moderation are not mutually exclusive phenomena. The validity of our statistical findings (e.g., the statistical moderating effects of classroom performance goal structures on the relations between personal performance-avoidance goals and student outcomes) does not rely on the assumption of the absence of mediation by personal goals. As explained by Muller, Judd, & Yzerbyt (2005), "the mediation question focuses on the intervening mechanism that produces the treatment effect. The moderation question focuses on factors that affect the magnitude of the treatment effect. It is important to note that these two processes may be combined in informative ways, such that moderation is mediated or mediation is moderated" (p. 852).

## References

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.
- Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, 84, 261–271.
- Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology*, 80, 260–267.
- Anderman, E. M., Maehr, M. L., & Midgley, C. (1999). Declining motivation after the transition to middle school: Schools can make a difference. *Journal of Research and Development in Education*, 32, 131–147.
- Anderman, E. M., & Midgley, C. (1997). Changes in personal achievement goals and the perceived classroom goal structures across the transition to middle level schools. *Contemporary Educational Psychology*, 22, 269–298.
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182.
- Barron, K. E., & Harackiewicz, J. M. (2001). Achievement goals and optimal motivation: Testing multiple goal models. *Journal of Personality and Social Psychology*, 80, 706–722.
- Bergman, L. R., Cairns, R. B., Nilsson, L., & Nystedt, L. (2000). *Developmental science and the holistic approach*. Mahwah, NJ: Erlbaum.
- Chan, D. (2006). Multilevel research. In F. T. L. Leong and J. T. Austin (Eds.), *The psychology research handbook: A guide for graduate students and research assistants*. Thousand Oaks, CA: Sage Publications.
- Church, M. A., Elliot, A. J., & Gable, S. L. (2001). Perceptions of classroom environment, achievement goals, and achievement outcomes. *Journal of Educational Psychology*, 93, 43–54.
- Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.
- Cronbach, L. J. (1976). *Research on classrooms and schools: Formulation of questions, design, and analysis* (Occasional Paper). Stanford, CA: Stanford Evaluation Consortium.
- Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist*, 41, 1040–1048.
- Dweck, C. S., & Leggett, E. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, 95, 256–273.
- Eccles, J. S., Midgley, C., Wigfield, A., Buchana, C. M., Reuman, D., Flanagan, C., et al. (1993). Development during adolescence: The impact of stage-environment fit on young adolescents' experiences in schools and in families. *American Psychologist*, 48, 90–101.
- Edwards, J. R., Cable, D. M., Williamson, I. O., Lambert, L. S., & Shipp, A. J. (2006). The phenomenology of fit: Linking the person and environment to the subjective experience of person-environment fit. *Journal of Applied Psychology*, 91, 802–827.
- Elliot, A. J. (2005). A conceptual history of the achievement goal structure. In A. J. Elliot, & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 52–72). New York: Guilford Press.
- Elliot, A. J., & Church, M. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology*, 72, 218–232.
- Elliot, A. J., & Harackiewicz, J. M. (1996). Approach and avoidance achievement goals and intrinsic motivation: A mediational analysis. *Journal of Personality and Social Psychology*, 70, 461–475.
- Elliot, A. J., & McGregor, H. A. (2001). A 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology*, 80, 957–971.
- Elliott, E. S., & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology*, 54, 5–12.
- Epstein, J. (1988). Effective schools or effective students: Dealing with diversity. In R. Haskins & D. MacRae (Eds.), *Policies for America's public schools: Teachers, equity, and indicators* (pp. 89–126). Norwood, NJ: Ablex.
- Glick, W. H. (1985). Conceptualizing and measuring organizational and psychological climate: Pitfalls in multilevel research. *Academy of Management Review*, 10, 601–616.
- Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Arnold.
- Harackiewicz, J. M., Barron, K. E., Pintrich, P. R., Elliot, A. J., & Thrash, T. M. (2002). Revision of achievement goal theory: Necessary and illuminating. *Journal of Educational Psychology*, 94, 638–645.
- Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environment* (3rd ed.). Odessa, FL: Psychological Assessment Resources.
- Hunt, D. E. (1975). Person-environment interaction: A challenge found wanting before it was tried. *Review of Educational Research*, 45, 209–230.
- Ingram, R. E., & Luxton, D. (2005). Vulnerability-stress models. In B. L. Hankin & J. R. Z. Abela (Eds.), *Development of psychopathology: A vulnerability-stress perspective* (pp. 32–46). New York: Sage.
- Ingram, R. E., & Price, J. M. (2001). *Vulnerability to psychopathology: Risk across the lifespan*. New York: Guilford Press.
- Kaplan, A., Gheen, M., & Midgley, C. (2002). Classroom goal structure and student disruptive behavior. *British Journal of Educational Psychology*, 72, 191–211.
- Kaplan, A., & Maehr, M. L. (1999). Achievement goals and student well-being. *Contemporary Educational Psychology*, 24, 330–358.
- Kaplan, A., & Middleton, M. J. (2002). Should childhood be a journal or a race? A response to Harackiewicz et al. (2002). *Journal of Educational Psychology*, 94, 646–648.
- Kaplan, A., Middleton, M. J., Urda, T., & Midgley, C. (2002). Achievement goals and goal structures. In C. Midgley (Ed.), *Goals, goal structures, and patterns of adaptive learning* (pp. 21–53). Mahwah, NJ: Erlbaum.
- Kaplan, A., & Midgley, C. (1999). The relationship between perceptions of the classroom goal structure and early adolescents' affect in school: The mediating role of coping strategies. *Learning and Individual Differences*, 11, 187–212.
- Karabenick, S. A. (2004). Perceived achievement goal structure and college student help seeking. *Journal of Educational Psychology*, 96, 569–581.
- Linnenbrink, E. A. (2004). Person and context: Theoretical and practical concerns in achievement goal theory. In P. R. Pintrich & M. L. Maehr (Eds.), *Advances in motivation and achievement: Motivating students, improving schools: The legacy of Carol Midgley* (Vol. 13, pp. 159–184). Greenwich, CT: Elsevier.



- Linnenbrink, E. A. (2005). The dilemma of performance-approach goals: The use of multiple goal contexts to promote students' motivation and learning. *Journal of Educational Psychology*, 97, 197-213.
- Linnenbrink, E. A., & Pintrich, P. R. (2001). Multiple goals, multiple contexts: The dynamic interplay between personal goals and contextual goal stresses. In S. Volet & S. Järvelä (Eds.), *Motivation in learning contexts: Theoretical and methodological implications* (pp. 251-269). Amsterdam: Pergamon Press.
- Maehr, M. L., & Midgley, C. (1996). *Transforming school cultures*. Boulder, CO: Westview Press.
- Magnusson, D., & Stattin, H. (1998). Person-context interaction theories. In W. Damon (Series Ed.) & R. M. Lerner (Vol. Ed.), *Handbook of child psychology: Vol. 1. Theoretical models of human development* (pp. 685-759). New York: Wiley.
- McClelland, G. H., & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, 114, 376-390.
- Meece, J. L., Anderman, E. M., & Anderman, L. H. (2006). *Annual Review of Psychology*, 57, 487-503.
- Meece, J. L., Blumenfeld, P. C., & Hoyle, R. H. (1988). Students' goal orientations and cognitive engagement in classroom activities. *Journal of Educational Psychology*, 80, 514-523.
- Midgley, C., Kaplan, A., & Middleton, M. (2001). Performance-approach goals: Good for what, for whom, under what circumstances, and at what cost? *Journal of Educational Psychology*, 93, 77-86.
- Midgley, C., Kaplan, A., Middleton, M., Maehr, M. L., Urdan, T., Anderman, L. H., et al. (1998). The development and validation of scales assessing students' achievement goal orientations. *Contemporary Educational Psychology*, 23, 113-131.
- Midgley, C., Maehr, M. L., Hruda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., et al. (2000). *Manual for the Patterns of Adaptive Learning Scales (PALS)*. Ann Arbor: University of Michigan.
- Mischel, W. (2003). *Introduction to personality* (7th ed.). New York: Wiley.
- Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102, 246-268.
- Monroe, S. M., & Simons, A. D. (1991). Diathesis-stress theories in the context of life stress research: Implications for the depressive disorders. *Psychological Bulletin*, 111, 406-425.
- Muller, D., Judd, C. M., & Yzerbyt, V. Y. (2005). When moderation is mediated and mediation is moderated. *Journal of Personality and Social Psychology*, 89, 852-863.
- Mullis, I. V. S., & Martin, M. O. (2000). Item analysis and review. In M. O. Martin, K. D. Gregory, & S. E. Stemler (Eds.), *TIMSS 1999 technical report* (pp. 225-234). Chestnut Hill, MA: Boston College, International Study Center.
- Newman, R. S. (1998). Students' help seeking during problem solving: Influences of personal and contextual achievement goals. *Journal of Educational Psychology*, 90, 644-658.
- Nicholls, J. G., Patashnick, M., & Nolen, S. B. (1985). Adolescents' theories of education. *Journal of Educational Psychology*, 77, 683-692.
- Nolen, S. B., & Haladyna, T. M. (1990). Motivation and studying in high school science. *Journal of Research in Science Teaching*, 27, 115-126.
- Patrick, H., Anderman, L. H., Ryan, A. M., Edelin, K., & Midgley, C. (2001). Teachers' communication of goal orientations in four fifth-grade classrooms. *Elementary School Journal*, 102, 35-58.
- Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). Orlando, FL: Harcourt Brace College Publishers.
- Pintrich, P. R. (2000). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology*, 92, 544-555.
- Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.
- Pintrich, P. R., Smith, D. A. F., Garcia, T., & McKeachie, W. J. (1993). Reliability and predictive validity of the motivated strategies for learning questionnaire (MSLQ). *Educational and Psychological Measurement*, 53, 801-813.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Roeser, R. W. (2004). Competing schools of thought in achievement goal theory? In P. R. Pintrich & M. L. Maehr (Eds.), *Advances in motivation and achievement: Motivating students, improving schools: The legacy of Carol Midgley* (Vol. 13, pp. 265-300). Greenwich, CT: Elsevier.
- Roeser, R. W., Midgley, C., & Urdan, T. C. (1996). Perceptions of the school psychological environment and early adolescents' psychological and behavioral functioning in school: The mediating role of goals and belonging. *Journal of Educational Psychology*, 88, 408-422.
- Rutter, M., Dunn, J., Plomin, R., Simonoff, E., Pickles, A., Maughan, B., et al. (1997). Integrating nature and nurture: Implications of person-environment correlations and interactions for developmental psychopathology. *Development and Psychopathology*, 9, 335-364.
- Ryan, A. M., Gheen, M., & Midgley, C. (1998). Why do some students avoid asking for help? An examination of the interplay among students' academic efficacy, teachers' socio-emotional role, and the classroom goal structure. *Journal of Educational Psychology*, 90, 528-535.
- Ryan, R. M., & Grolnick, W. S. (1986). Origins and pawns in the classroom: Self-report and projective assessments of individual differences in children's perceptions. *Journal of Personality and Social Psychology*, 50, 550-558.
- Snijders, T. B., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.
- Steinberg, L., Lamborn, S. D., Dornbusch, S. M., & Darling, N. (1992). Impact of parenting practices on adolescent achievement: Authoritative parenting, school involvement, and encouragement to succeed. *Child Development*, 63, 1266-1281.
- Turner, J. C., Midgley, C., Meyer, D. K., Gheen, M., Anderman, E. M., Kang, Y., et al. (2002). The classroom environment and students' reports of avoidance strategies in mathematics: A multimethod study. *Journal of Educational Psychology*, 94, 88-106.
- Urdan, T. (1997). Achievement goal theory: Past results, future directions. In M. L. Maehr & P. R. Pintrich (Eds.), *Advances in motivation and achievement* (Vol. 10, pp. 99-141). Greenwich, CT: JAI Press.
- Urdan, T. (2004). Can achievement goal theory guide school reform? In P. R. Pintrich & M. L. Maehr (Eds.), *Advances in motivation and achievement: Motivating students, improving schools: The legacy of Carol Midgley* (Vol. 13, pp. 361-392). Greenwich, CT: Elsevier.
- Urdan, T., Midgley, C., & Anderman, E. M. (1998). The role of classroom goal structure in students' use of self-handicapping. *American Educational Research Journal*, 35, 101-122.
- Urdan, T., Ryan, A. M., Anderman, E. M., & Gheen, M. H. (2002). Goals, goal structures, and avoidance behaviors. In C. Midgley (Ed.), *Goals, goal structures, and patterns of adaptive learning* (pp. 55-83). Mahwah, NJ: Erlbaum.
- Wachs, T. D. (2000). *Necessary but not sufficient: The role of individual and multiple influences on human development*. Washington, D.C.: American Psychological Association.
- Wellborn, J. G., & Connell, J. P. (1987). *Manual for the Rochester Assessment Package for Schools*. Rochester, NY: University of Rochester.
- Wolters, C. A. (2004). Advancing achievement goal theory: Using goal structures and goal orientations to predict students' motivation, cognition, and achievement. *Journal of Educational Psychology*, 96, 236-250.
- Wolters, C. A., Yu, S. L., & Pintrich, P. R. (1996). The relation between goal orientation and students' motivational beliefs and self-regulated learning. *Learning and Individual Differences*, 8, 211-238.

## Appendix

## Sample Items for Self-Report Scales

*Classroom Mastery Goal Structure (4 items)*

My math teacher wants us to really understand the subject, not just to remember facts or rules.

*Classroom Performance Goal Structure (5 items)*

My math teacher thinks that it is more important to do well in math tests than to learn new things.

*Personal Mastery Goal (5 items)*

An important reason I do my math work is that I like to learn new things.

*Personal Performance-Approach Goal (4 items)*

I like to show my teacher that I am smarter than the other pupils in my math class.

*Personal Performance-Avoidance Goal (4 items)*

I do my math work because I do not want the teacher to think that I am stupid.

*Interest (4 items)*

I enjoy doing math.

*Engagement (5 items)*

I pay attention well in my math class.  
I try my best to contribute during small group discussions.

*Effort Withdrawal (4 items)*

I do not work hard on my math homework.

*Avoidance Coping (3 items)*

When the work in math is difficult, I give up.

Received August 10, 2006

Revision received April 24, 2007

Accepted June 23, 2007 ■



# Problem Solving and Computational Skill: Are They Shared or Distinct Aspects of Mathematical Cognition?

Lynn S. Fuchs and Douglas Fuchs  
Vanderbilt University

Karla Stuebing and Jack M. Fletcher  
University of Houston

Carol L. Hamlett and Warren Lambert  
Vanderbilt University

The purpose of this study was to explore patterns of difficulty in 2 domains of mathematical cognition: computation and problem solving. Third graders ( $n = 924$ ; 47.3% male) were representatively sampled from 89 classrooms; assessed on computation and problem solving; classified as having difficulty with computation, problem solving, both domains, or neither domain; and measured on 9 cognitive dimensions. Difficulty occurred across domains with the same prevalence as difficulty with a single domain; specific difficulty was distributed similarly across domains. Multivariate profile analysis on cognitive dimensions and chi-square tests on demographics showed that specific computational difficulty was associated with strength in language and weaknesses in attentive behavior and processing speed; problem-solving difficulty was associated with deficient language as well as race and poverty. Implications for understanding mathematics competence and for the identification and treatment of mathematics difficulties are discussed.

**Keywords:** calculations, word problems, cognitive predictors, mathematics

Mathematics, which involves the study of quantities as expressed in numbers or symbols, comprises a variety of related branches. In elementary school, for example, mathematics is conceptualized in strands such as concepts, numeration, measurement, arithmetic, algorithmic computation, and problem solving. In high school, curriculum offerings include algebra, geometry, trigonometry, and calculus. Little is understood, however, about how different aspects of mathematical cognition relate to one another (i.e., which aspects of performance are shared or distinct, or how difficulty in one domain corresponds with difficulty in another). Such understanding would provide theoretical insight into the nature of mathematics competence and practical guidance about the identification and treatment of mathematics difficulties.

The purpose of the present study was to explore the overlap of difficulty with two aspects of primary-grade mathematical cognition and to examine how characteristics differ among subgroups with difficulty in one, the other, both, or neither. The first aspect of performance was computation, including skill with number

combinations (e.g.,  $2 + 5$ ;  $8 - 3$ ) and procedural computation (e.g.,  $25 + 38$ ;  $74 - 22$ ). The second aspect of performance was problem solving, including one-step, contextually straightforward word problems (e.g., John had 9 pennies. He spent 3 pennies at the store. How many pennies did he have left?) and multistep, contextually more complex problems (e.g., Fred went to the ballgame with 2 friends. He left his house with \$42. While at the game, he bought 5 hotdogs and 3 sodas. The hot dogs cost \$10 each, and the sodas cost \$5 each. How much did Fred spend?).

The major distinction between computation and problem solving is the addition of linguistic information that requires children to construct a problem model. Whereas a computation problem is already set up for solution, a word problem requires students to use the text to identify missing information, construct the number sentence, and derive the calculation problem for finding the missing information. This transparent difference would seem to alter the nature of the task, but no studies have examined how difficulty in one subdomain corresponds to difficulty in the other and whether students' cognitive characteristics differ as a function of where the mathematics difficulty resides.

By contrast, a related literature does focus on the interplay between math and reading. In these studies, math difficulty typically is defined on a broad measure tapping multiple aspects of performance, then subgroups are formed to determine how performance on mathematics domains differs with or without concurrent reading difficulty. Research has shown that students with difficulty in both math and reading (usually defined in terms of word recognition) experience more pervasive deficits in computation (e.g., Jordan & Hanich, 2000) and problem solving (e.g., Fuchs & Fuchs, 2002; Hanich, Jordan, Kaplan, & Dick, 2001; Jordan & Hanich, 2000). This may occur due to a different pattern of

---

Lynn S. Fuchs, Douglas Fuchs, and Carol L. Hamlett, Department of Special Education, Vanderbilt University; Karla Stuebing and Jack M. Fletcher, Department of Psychology, University of Houston; Warren Lambert, Kennedy Center for Research on Human Development, Vanderbilt University.

This research was supported in part by Grant 1 RO1 HD46154 and Core Grant HD15052 from the National Institute of Child Health and Human Development to Vanderbilt University. Statements do not reflect agency position or policy, and no official endorsement should be inferred.

Correspondence concerning this article should be addressed to Lynn S. Fuchs, Peabody College, Box 228, Vanderbilt University, Nashville, TN 37203. E-mail: lynn.fuchs@vanderbilt.edu

underlying deficits in domain-general abilities associated with comorbidity. Other work (e.g., Jordan, Hanich, & Kaplan, 2003; Landerl, Began, & Butterworth, 2004), however, has shown that students with general math difficulty, with and without reading problems, experience comparable deficits on number combinations. An older body of research suggested that specific math difficulty, usually defined as performance on a broad computational task, is associated with difficulties in nonverbal processing (spatial cognition, working memory) and procedural knowledge (Geary, 1993; Rourke & Finlayson, 1978); concurrent reading and math difficulties reflect more pervasive language and working-memory problems.

This line of work, which speaks to the issue of reading difficulty, is important for generating hypotheses about the nature of mathematics disability as well as its identification and treatment. This literature does not, however, address the issue of whether difficulty within mathematics domains is better conceptualized as shared or distinct. Four large-scale studies speak indirectly to this issue by examining the cognitive characteristics associated with computational and problem-solving skill among representative samples. Studying 353 first through third graders, Swanson and Beebe-Frankenberger (2004) identified working memory as an ability that contributed to strong performance across both areas of mathematical cognition, but some unique cognitive abilities also emerged as important: phonological processing for computation and fluid intelligence as well as short-term memory for simple word problems. In an extension of this work following students' development of calculations and problem-solving skill over 1 year, Swanson (2006) identified predictors of computation (inhibition or controlled attention, vocabulary knowledge, visual-spatial working memory) that differed from problem solving (working-memory's executive system, operationalized as listening span, backward digit span, and digit/sentence span). A latent variable of reading (i.e., phonological processing, timed and untimed reading of real words, timed reading of pseudowords, and comprehension) accounted for skill in both math outcomes.

With a sample of 312 third graders, Fuchs et al. (2006) examined the concurrent cognitive correlates of computation versus simple word problems, this time controlling for the role of arithmetic skill within simple word problems. Teacher ratings of inattentive behavior were identified as a correlate common to both subdomains of math, but the remaining abilities differed: for computation, phonological decoding and processing speed; for word problems, nonverbal problem solving, concept formation, sight word efficiency, and language. The fourth study (Fuchs et al., 2005) used beginning-of-the-year cognitive abilities to predict the development of skill across the year among 272 first graders. Results again suggested some common and some unique patterns of cognitive abilities. The common predictors were working memory and ratings of attention. The unique predictors were phonological processing for computation and nonverbal problem solving for simple word problems.

Across these studies, some findings recur; others are idiosyncratic. But together, the results indicate that some abilities underlying these domains of mathematical cognition are unique. This provides the basis for hypothesizing that the cognitive dimensions underlying difficulty in each of these domains may also be distinct and that difficulty in these two domains may be distinct. This has implications for understanding, identifying, and treating mathe-

matics difficulties. In a related way, such knowledge would help determine whether the distinction in these domains newly introduced into the 2004 reauthorization of the Individuals with Disabilities Education Act is viable. That is, although the reauthorized law, which guides the identification and treatment of students with disabilities throughout schools in the United States, makes a distinction between problem-solving and computational forms of mathematics disability, little prior work is available to assess the validity of this distinction.

In the present study, we extended the literature by addressing these issues directly. We began by conducting preliminary analyses to evaluate the predictors of each math outcome using assessments of math skills and cognitive domains in a large population-based sample of children in third grade. In line with the four earlier studies on representative samples, we hypothesized that different cognitive skills would be associated with each domain, even when we accounted for shared variance in computation and problem solving. Our major analyses, however, focused specifically on determining whether children with extreme deficits in computation or problem solving represent distinct groups. We identified students at the lower end of the distribution on computation, on problem solving, on both, or on neither, and we examined the extent to which students actually experienced difficulty in one domain but not the other. Then, we assessed how the demographic and cognitive profiles associated with these subgroups differed. Using profile analysis, we hypothesized that groups based on computation or problem solving would show different profiles and that the presence of both difficulties would show features of both domains, reflecting a comorbid association. We contrasted profile analysis findings based on a univariate versus a multivariate approach.

We note that the vast majority of prior work examining the cognitive correlates of primary-grade mathematics performance focused on a limited set of cognitive abilities related to a single aspect of math skill, rather than studying how these abilities operate within a multivariate framework to explain different aspects of mathematical cognition. For this reason, the literature provides the basis for deliberate hypotheses about which cognitive abilities may mediate which aspect of third-grade math performance. The literature does not, however, provide the basis for specifying an integrated theory about how these variables operate in coordinated fashion to explain computation versus problem solving. Before describing the method of the present study, we summarize the basis for our hypotheses about which cognitive dimensions might be related to which aspect of mathematical cognition.

In terms of computation, prior work provides the basis for hypothesizing that attentive behavior, working memory, and processing speed may help determine skill. Because computation requires a series of steps, attentive behavior (i.e., low distractibility) may enhance performance. Russell and Ginsburg (1984) provided suggestive evidence on this possibility when they compared math-disabled fourth graders to normal fourth graders and normal third graders. Results indicated that the algorithmic errors of math-disabled students were similar to those of both normal groups, but math-disabled students more closely resembled younger normal counterparts in detecting those errors. More recently, Swanson (2006) showed that inhibitory control predicted the development of computation but not problem-solving skill.



Second, prior work has implicated *working memory* (e.g., Geary, Brown, & Samaranayake, 1991; Hitch & McAuley, 1991; Siegel & Linder, 1984; Webster, 1979; Wilson & Swanson, 2001), or the capacity to maintain target memory items while processing an additional task (Daneman & Carpenter, 1980). Although the relation between working memory and memory-based retrieval of computation has been repeatedly documented, the nature of that relation is unclear. As described by Geary (1993), working memory involves component skills including, but not limited to, rate of decay (creating difficulties in holding the association between a problem stem and its answer) and attentive behavior (hence the finding that math-disabled children monitor problem solving less well than non-math-disabled children; Butterfield & Ferretti, 1987; Geary, Widaman, Little, & Cormier, 1987). In addition, memory span appears to be related to how quickly numbers can be counted (Geary, 1993).

It is not surprising, therefore, that *processing speed*, or the efficiency with which simple cognitive tasks are executed (R. Case, 1985), represents a promising candidate. Processing speed may dictate how quickly numbers can be counted. With slower processing, the interval for deriving counted answers and for pairing a problem stem with its answer in working memory increases; this creates the possibility that decay sets in before completing the computational sequence. Bull and Johnston (1997) found that processing speed was the best predictor of computational competence among 7-year-olds, subsuming all of the variance accounted for by long- and short-term memory, even with reading performance controlled. More recently, Hecht, Torgesen, Wagner, and Rashotte (2001) provided corroborating data on the importance of processing speed as a correlate of computational skill while controlling for vocabulary knowledge.

As for problem solving, prior work examining which cognitive processes mediate arithmetic word problems has focused heavily on working memory, probably because research (e.g., Hitch & McAuley, 1991; Siegel & Ryan, 1989) shows that children with learning disabilities experience concurrent difficulty with working memory (e.g., Siegel & Ryan, 1989; Swanson, Ashbaker, & Sachse-Lee, 1996) and mathematical problem solving (e.g., L. P. Case, Harris, & Graham, 1992; Swanson, 1993). Also, theoretical frameworks (e.g., Kintsch & Greeno, 1985; Mayer, 1992) posit that word problems involve construction of a problem model, which appears to require working-memory capacity. For example, according to Kintsch and Greeno, when people solve word problems, new sets are formed on-line as the story is processed. When a proposition that triggers a set-building strategy is completed, the appropriate set is formed and the relevant propositions are assigned places in the schema. As new sets are formed, previous sets that had been active in the memory buffer are displaced.

In line with theoretical models implicating working memory, the literature provides support for its importance. For example, Passolunghi and Siegel (2001) found that 9-year-olds, characterized as good or poor problem solvers, differed on working-memory tasks. Other researchers have found corroborating evidence using similar methods (e.g., LeBlanc & Weber-Russell, 1996; Passolunghi & Siegel, 2004; Swanson & Sachse-Lee, 2001). At the same time, other studies have raised questions about the robustness of the relation. For example, among typically developing third and fourth graders, Swanson, Cooney, and Brock (1993) found only a weak relation between working memory and problem solution accuracy,

and this relation disappeared once reading comprehension was considered. The other leading candidates are attentive behavior, nonverbal problem solving, language ability, reading skill, and concept formation.

In studies involving attention, most work has focused on the inhibition of irrelevant stimuli, with mixed results. Passolunghi and colleagues ran a series of studies suggesting the importance of inhibition. For example, comparing good and poor problem solvers, Passolunghi, Cornoldi, and De Liberto (1999) found comparable storage capacity with inefficiencies of inhibition (i.e., poor problem solvers remembered less relevant but more irrelevant information in math problems). In contrast, Swanson and Beebe-Frankenberger (2004) and Swanson (2006) found no evidence that inhibition contributed to problem-solving skill. Research has, however, rarely studied the role of attention more broadly. An exception is Fuchs et al. (2005), who found that a teacher rating scale of attentive behavior predicted the development of first-grade skill with word problems.

*Nonverbal problem solving*, or the ability to complete patterns presented visually, has been identified as a unique predictor in the development of problem-solving skill across first grade (Fuchs et al., 2005), a finding corroborated by Agness and McLone (1987). This is not surprising, because word problems, in which the problem narrative poses a question entailing relationships between numbers, appear to require conceptual representations. Language ability is also important to consider given the obvious need to process linguistic information when building a problem representation of an arithmetic word problem. In fact, Jordan, Levine, and Huttenlocher (1995) documented the importance of language ability when they showed that kindergarten and first-grade language-impaired children (receptive vocabulary and grammatic closure < 30th percentile) performed significantly lower than nonimpaired peers on arithmetic word problems. Finally, it is hard to ignore the possibility that reading skill may underlie skill in problem solving. Reading is transparently involved, even when problems are read aloud to children, because reading skill provides continuing access to the written problem narrative after the adult reading has been completed. This potentially reduces the load on working memory and thereby facilitates solution accuracy. Swanson (2006) recently identified reading as a predictor of computational as well as problem-solving skill.

## Method

### Participants

The data described in this paper were collected as part of a prospective 4-year study assessing the effects of mathematical problem-solving instruction and examining the developmental course and cognitive predictors of mathematical problem solving. The data in the present article were collected with the first-, second-, and third-year cohorts at the first assessment wave, sampling from 1,958 students in 89 third-grade classrooms in 10 Title 1 schools and 3 non-Title 1 schools in a southeastern metropolitan school district.

The sampling process was designed to yield a representative sample. That is, from these 1,958 students, we randomly sampled 990 students for participation, blocking within classroom and within three strata: (a) 25% of students with scores 1 *SD* below the



mean of the entire distribution on the Test of Computational Fluency (see *Math Measures*); (b) 50% of students with scores within 1 *SD* of the mean of the entire distribution on the Test of Computational Fluency; and (c) 25% of students with scores 1 *SD* above the mean of the entire distribution on the Test of Computational Fluency. Of these 990 students, we had complete data for the variables reported in the present study on 924 children, who were the basis for the present report. As measured on the two-subtest Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999), IQ averaged 97.68 (*SD* = 14.26). Normal curve equivalent scores on the TerraNova (CTB/McGraw-Hill, 1997), administered the previous spring by the school district, averaged 55.40 (*SD* = 16.72) for the reading composite and 55.75 (*SD* = 20.15) for the mathematics composite. Standard scores on the Woodcock-Johnson III Tests of Achievement (WJ-III; Woodcock, McGrew, & Mather, 2001) Applied Problems averaged 102.25 (*SD* = 13.51), and standard scores on the Woodcock Reading Mastery Tests—Revised Word Identification (Woodcock, 1998) averaged 101.05 (*SD* = 10.05). Of the 924 students, 437 (47.3%) were male, and 499 (54.0%) received subsidized lunch. Ethnicity was distributed as follows: 395 (42.5%) African American, 363 (39.3%) European American, 94 (10.2%) Hispanic, 17 (1.8%) Kurdish, and 55 (6.0%) other. Schools had identified 73 students (7.9%) as having a disability (i.e., learning disability, speech impairment, language impairment, attention-deficit/hyperactivity disorder, health impairment, or emotional behavioral disorder).

### Procedure

In this report, we describe only the subset of measures on which we report data. The math measures were administered in large-group arrangement in September of third grade during three sessions each lasting 30 to 60 min. These large-group sessions included three tests of computational skill (Addition Fact Fluency, Subtraction Fact Fluency, and Test of Computational Fluency) and three tests of word problem skill (Simple Word Problems, Algorithmic Word Problems, and Complex Word Problems). Measures of eight (of the nine) cognitive dimensions were administered individually in September and October of third grade during two 45-min sessions: Woodcock Diagnostic Reading Battery Listening Comprehension, Test of Language Development—Primary Grammatical Closure, WASI Vocabulary, WJ-III Retrieval Fluency, WJ-III Concept Formation, WASI Matrix Reasoning, Working Memory Test Battery for Children Listening Recall, WJ-III Numbers Reversed, Woodcock Reading Mastery Tests—Revised Word Identification, and WJ-III Visual Matching. Tests were administered by trained examiners, each of whom had demonstrated 100% accuracy during mock administrations. All individual sessions were audiotaped, and 19.7% of tapes, distributed equally across testers, were selected randomly for accuracy checks by an independent scorer. Agreement was between 98.7% and 99.9%. In October, classroom teachers completed the Strengths and Weaknesses of Attention-Deficit/Hyperactivity Disorder-Symptoms and Normal-Behavior (SWAN) Rating Scale, the ninth cognitive dimension, on each student.

### Math Measures

*Addition and subtraction fact fluency.* The Grade 3 Math Battery (Fuchs, Hamlett, & Powell, 2003) incorporates two math

fact retrieval subtests. Addition Fact Fluency comprises 25 addition fact problems with answers from 0 to 12 and with addends from 0 to 9. Problems are presented horizontally on one page. Students have 1 min to write answers. The score is the number of correct answers. Agreement, calculated on 20% of protocols by two independent scorers, was 99.8%. For the representative sample, coefficient alpha was .91; criterion validity with the previous spring's TerraNova (CTB/McGraw-Hill, 1997) Total Math score was .53 for the 844 students for whom we had TerraNova scores. Subtraction Fact Fluency comprises 25 subtraction fact problems with answers from 0 to 12 and with minuends/subtrahends from 0 to 18. Problems are presented horizontally on one page. Students have 1 min to write answers. The score is the number of correct answers. Agreement, calculated on 20% of protocols by two independent scorers, was 98.5%. For the representative sample, coefficient alpha was .92, and criterion validity with the previous spring's TerraNova Total Math score was .51 for the 844 students for whom we had TerraNova scores.

*Procedural computation.* The Test of Computational Fluency (Fuchs, Hamlett, & Fuchs, 1990) is a one-page test displaying 25 items that sample the typical second-grade computation curriculum, including adding and subtracting number combinations and algorithmic computation. Students have 3 min to complete as many answers as possible. The score is the number of correct responses. Staff entered responses into a computerized scoring program on an item-by-item basis, with 15% of tests reentered by an independent scorer. Data-entry agreement was 99.6%. For the representative sample, coefficient alpha was .94, and criterion validity with the previous spring's TerraNova (CTB/McGraw-Hill, 1997) Total Math score was .60 for the 844 students for whom we had TerraNova scores.

*Simple arithmetic word problems.* Following Jordan and Hanich (2000; adapted from Carpenter & Moser, 1984; Riley & Greeno, 1988; Riley, Greeno, & Heller, 1983), Story Problems comprises 14 one-step word problems that express change, combine, compare, and equalize relationships among numbers and require sums or subtrahends of 9 or less for solution. The tester reads each item aloud while students follow along on their own copies of the problems. Students have 30 s to respond to each item before the tester moves to the next one, and students can ask for rereading(s) as needed. The score is the number of correct answers. A second scorer independently rescored 20% of protocols, with agreement of 99.9%. For the representative sample, coefficient alpha was .86, and criterion validity with the previous spring's TerraNova (CTB/McGraw-Hill, 1997) Total Math score was .62 for the 844 students for whom we had TerraNova scores.

*Algorithmic word problems.* Algorithmic Word Problems (Fuchs et al., 2003) comprises 10 word problems, each of which requires one to four steps. The measure samples four problem types, asking students to (a) apply step-up functions, (b) add multiple quantities of items each with different prices, (c) find half, or (d) sum a quantity derived from a pictograph with another addend. The tester reads each item aloud while students follow along on their own copies of the problems. The tester progresses to the next problem when all but one or two students have their pencils down, indicating they are finished. Students can ask for rereading(s) as needed. The maximum score is 44. For the representative sample, Cronbach's alpha was .85, and criterion validity with the previous spring's TerraNova (CTB/McGraw-Hill, 1997)



Total Math score was .58 for the 844 students for whom we had TerraNova scores. Interscorer agreement, computed on 20% of protocols by two independent scorers, was .984.

*Complex word problems.* Complex Word Problems (Fuchs et al., 2003) comprises nine problems representing the same four problem types as the algorithmic word problems within more complex contexts: (a) adding multiple quantities of items with different prices, with information presented in bulleted format and with a selection response format; (b) adding multiple quantities of items with different prices, also asking for money left at the end; (c) a step-up function problem with irrelevant information; (d) a step-up function that requires students to compare the prices of two packaging options; (e) a half problem using the words *share equally* instead of *half*; (f) a pictograph/adding problem asking for money left at the end; (g) a pictograph/adding problem comparing two quantities; (h) a problem with irrelevant information that combines multiple quantities with different prices and pictograph/adding; and (i) a problem with irrelevant information that combines multiple quantities with different prices and a step-up function. The tester reads each item aloud while students follow along on their own copies of the problems. The tester progresses to the next item when all but one or two students have their pencils down, indicating they are finished. Students can ask for rereading(s) as needed. The maximum score is 79. For the representative sample, Cronbach's alpha was .88, and criterion validity with the previous spring's TerraNova (CTB/McGraw-Hill, 1997) Total Math Score was .55 for the 844 students for whom we had TerraNova scores. Interscorer agreement, computed on 20% of protocols by two independent, blind scorers, was .983.

### Cognitive Dimensions

*Language.* Using three measures of language skill, we used a principal components factor analysis to create a weighted composite variable of language. Test of Language Development—Primary Grammatical Closure (Newcomer & Hammill, 1988) measures the ability to recognize, understand, and use English morphological forms. The examiner reads 30 sentences, one at a time. Each sentence has a missing word, and examinees earn 1 point for each sentence correctly completed. As reported by the test developers, reliability is .88 for 8-year-olds; the correlation with the Illinois Test of Psycholinguistic Ability Grammatical Closure is .88 for 8-year-olds. Coefficient alpha on the representative sample was .76. The Woodcock Diagnostic Reading Battery Listening Comprehension (Woodcock, 1997) measures the ability to understand sentences or passages. For 38 items, students supply the word missing from the end of each sentence or passage. The test begins with simple verbal analogies and associations and progresses to comprehension involving the ability to discern implications. Testing is discontinued after six consecutive errors. The score is the number of correct responses. As reported by the test developers, reliability is .80 at ages 5 to 18; the correlation with the Woodcock-Johnson Psycho-Educational Battery—Revised (Woodcock & Johnson, 1989) is .73. Coefficient alpha on the representative sample was .81. WASI Vocabulary (Wechsler, 1999) measures expressive vocabulary, verbal knowledge, and foundation of information with 42 items. The first four items present pictures; the student identifies the object in the picture. For remaining items, the tester says a word that the student defines.

Responses are awarded a score 0, 1, or 2 depending on quality. Testing is discontinued after five consecutive scores of 0. The score is the total number of points. As reported by Zhu (1999), split-half reliability is .86 to .87 at ages 6 to 7; the correlation with the Wechsler Intelligence Scale for Children—III Full Scale IQ is .72. Coefficient alpha on the representative sample was .78.

*Semantic retrieval fluency.* WJ-III Retrieval Fluency (Woodcock et al., 2001) asks examinees to recall related items, within categories, for 1 min per category. Examinees earn credit for each nonduplicated answer. As reported by the test developers, reliability is .78 for 8-year-olds.

*Concept formation.* WJ-III Concept Formation (Woodcock et al., 2001) asks examinees to identify the rules for concepts when shown illustrations of instances and noninstances of the concept. Examinees earn credit by correctly identifying the rule that governs each concept. Cutoff points determine the ceiling. The score is the number of correct responses. As reported by the test developers, reliability is .93 for 8-year-olds. Coefficient alpha on the representative sample was .82.

*Nonverbal problem solving.* WASI Matrix Reasoning (Wechsler, 1999) measures nonverbal reasoning with four types of tasks: pattern completion, classification, analogy, and serial reasoning. Examinees look at a matrix from which a section is missing and complete the matrix by saying the number of or pointing to one of five response options. Examinees earn points by identifying the correct missing piece of the matrix. Testing is discontinued after four errors on five consecutive items or after four consecutive errors. The score is the number of correct responses. As reported by the test developer, reliability is .94 for 8-year-olds; the correlation with the Wechsler Intelligence Scale for Children—III Full Scale IQ is .66. Coefficient alpha on the representative sample was .76.

*Working memory.* With the Working Memory Test Battery for Children Listening Recall (Pickering & Gathercole, 2001), a measure of verbal working memory, the tester says a series of short sentences, only some of which make sense. The student indicates whether each sentence is true or false. After all sentences in a trial (i.e., one to six sentences) are heard and determined to be true or false, the student recalls the final word of each sentence in the order presented. The student earns 1 point for each sequence of final words recalled correctly in the right order, and the score is the total of correct sequences. Testing is discontinued when the student makes three or more errors in any block of items. As reported by Pickering and Gathercole, test-retest reliability is .93. Coefficient alpha on the representative sample was .72. With WJ-III Numbers Reversed (Woodcock et al., 2001), a measure of numerical working memory, the tester says a string of random numbers; the student says the series backwards. Item difficulty increases as more numbers are added to the series. Students earn credit by repeating the numbers correctly in the opposite order. As reported by the test developers, reliability is .86 for 8-year-olds. Coefficient alpha on the representative sample was .84.

*Word identification skill.* The Woodcock Reading Mastery Tests—Revised Word Identification (Woodcock, 1998) measures real-word reading ability with 100 words arranged in order of difficulty. Students read words aloud. Testing is discontinued after six consecutive errors at the end of a page. The score is the number of correct items. As reported by Woodcock, split-half reliability is .98. Coefficient alpha on the representative sample was .87.



**Attentive behavior.** The SWAN is an 18-item teacher rating scale (www.adhd.net). Items from the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; American Psychiatric Association, 1994) criteria for attention-deficit/hyperactivity disorder are included for inattention (largely distractibility; Items 1–9) and hyperactivity/impulsivity (Items 10–18). Items are rated on a scale of 1 to 7 (1 = *far below*, 2 = *below*, 3 = *slightly below*, 4 = *average*, 5 = *slightly above*, 6 = *above*, 7 = *far above*). In the present study, we report data for the inattentive behavior subscale. Using the nine relevant items, we used a principal components factor analysis to create a weighted composite variable of attentive behavior, or the ability to maintain focus. (Because the principal components factor analysis yielded only one factor, no rotation was necessary.) The SWAN has been shown to correlate well with other dimensional assessments of behavior related to inattention (www.adhd.net). Coefficient alpha on the representative sample was .96.

**Processing speed.** WJ-III Visual Matching (Woodcock et al., 2001) measures processing speed by asking examinees to locate and circle two identical numbers in rows of six numbers. Examinees have 3 min to complete 60 rows and earn credit by correctly circling the matching numbers in each row. As reported by the test developer, reliability is .91 for 8-year-olds.

## Data Analysis and Results

### Variable Transformations

In Table 1, we show means, standard deviations, and correlations for the representative sample of 924 students on computation, problem solving, and nine cognitive dimensions (language, semantic retrieval fluency, concept formation, matrix reasoning, verbal working memory, numerical working memory, word identification, attentive behavior, and processing speed). Based on the entire sample of 924 students, we transformed raw scores for each of the three computation and each of the four problem solving measures to *z* scores ( $M = 0.00$ ,  $SD = 1.00$ ). For dimensions with more than one indicator (computation, problem solving, language, and attention), we used SAS PROC FACTOR to estimate factor scores, using squared multiple correlations as the communalities.

### Preliminary Analyses: Screening for Outliers and Nonlinearity and Regressions on the Representative Sample

To identify potential outliers, we plotted the bivariate relations of both problem solving and computation with each cognitive variable. We identified five outlier values whose cases we eliminated from further analyses. Then, to investigate the shape of the relations, we examined the bivariate relations between each math outcome and each cognitive variable. We used linear and quadratic forms of each predictor to investigate the functional relation. Any significant quadratic relations between cognitive dimension and math outcome were marked for inclusion in the full regression analysis. We found significant (although not substantial) nonlinearity between problem solving and computation when predicting computation from problem solving (but not when predicting problem solving from computation). So, in the full regression predict-

ing computation from problem solving, we included the quadratic term for problem solving.

We then completed regression analyses to examine the relation of each math outcome with the nine cognitive dimensions using the entire sample. When the quadratic relation between a cognitive dimension and a math outcome was significant (see above), we retained both the linear and quadratic relations within the full model. (We note that significant nonlinear relations between the cognitive variables and the math outcomes were trivial.) For each math outcome, we also investigated the interaction of the cognitive variables with the math outcome variable we were controlling. For example, when predicting computation, we entered problem solving, all nine cognitive variables, any significant quadratic relations (including the quadratic relation for problem solving), and the product vectors of each cognitive variable with problem solving. We did this to determine whether the relation of the cognitive variables with computation was consistent over the range of problem solving. We found that the block of interaction vectors did not add significantly for either math outcome and deleted the product vectors from the model. Finally, for each math outcome, we ran a pruned model, the variables for which neither the linear nor quadratic relation in the full regression model was significant. In Table 2, we present the results of the full and pruned models predicting computation performance. In Table 3, we present findings for the prediction of problem-solving performance. As shown, after we controlled for problem-solving skill in the prediction of computational skill, significant cognitive predictors were word identification, attentive behavior, processing speed, as well as the quadratic relation for numerical working memory. By contrast, after we controlled for computational skill in the prediction of problem-solving skill, significant cognitive predictors were language, concept formation, matrix reasoning, numerical working memory, as well as the quadratic relation for language and attentive behavior.

### Difficulty Status Group Formation

Because the major focus of the present study was to extend understanding about difficulty in computation versus problem solving and because the regressions only informed about the effect of each variable when the other variables are controlled at their means, it was important to examine performance specifically at the lower ranges of performance on the math outcomes. To establish extreme groups impaired on computation, problem solving, or both, we designated math difficulty in the following manner. Any student who scored above the 40th percentile on the problem-solving factor score and above the 40th percentile on the computation factor score was designated as having no difficulty (ND). Any student who scored below the 15th percentile on the computation factor score but above the 40th percentile on the problem-solving factor score was designated as having computational difficulty (CD). Any student who scored below the 15th percentile on the problem-solving factor score but above the 40th percentile on the computation factor score was designated as having problem-solving difficulty (PD). Any student who scored below the 15th percentile on the problem-solving factor score and below the 15th percentile on the computation factor score was designated as having computation and problem-solving difficulty (CPD). This placed 372 students in the buffer zone (i.e., scoring between the 16th and 39th percentiles on either or both math outcome) and



Table 1  
Means, Standard Deviations, and Correlations on Representative Sample ( $n = 924$ )

Means, Standard Deviations, and Correlations on Representative Sample (*n* = 924)

Variable	Raw		Transformed		Correlations															
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1. Basic facts <sup>+a</sup>	12.11	4.99	0.00	1.00	—															
2. Basic facts <sup>—b</sup>	7.00	5.07	0.00	1.00	.59	—														
3. Computation <sup>c</sup>	12.37	6.15	0.00	1.00	.71	.68	—													
4. Simple PS <sup>d</sup>	10.01	3.44	0.00	1.00	.39	.40	.47	—												
5. Algorithmic PS <sup>e</sup>	7.70	5.99	0.00	1.00	.40	.44	.49	.54	—											
6. Complex PS <sup>f</sup>	6.13	6.05	0.00	1.00	.33	.37	.45	.46	.63	—										
7. Real-world PS	9.47	9.88	0.00	1.00	.30	.35	.41	.43	.54	.47	—									
8. Language <sup>g</sup>			0.00	1.00	.23	.26	.33	.53	.46	.43	.47	—								
Grammatic	18.76	6.74	85.59	11.21																
Listening	21.05	4.35	96.48	18.27																
Vocabulary	27.89	6.53	47.24	10.14																
9. Semantic retrieval <sup>h</sup>	496.77	3.64	93.23	14.07	.24	.21	.25	.27	.27	.23	.25	.42	—							
10. Concept formation <sup>i</sup>	15.90	7.19	93.04	13.60	.29	.30	.35	.48	.45	.44	.44	.51	.27	—						
11. Matrix reasoning <sup>j</sup>	15.85	6.41	49.21	11.10	.21	.26	.29	.39	.40	.37	.35	.36	.20	.38	—					
12. Verbal working memory <sup>k</sup>	9.91	3.22	92.37	17.45	.22	.24	.24	.38	.34	.32	.31	.48	.23	.41	.30	—				
13. Numerical working memory <sup>l</sup>	9.39	2.80	95.89	14.05	.23	.28	.24	.32	.33	.26	.25	.27	.20	.29	.32	.36	—			
14. Word ID <sup>m</sup>	57.17	9.87	101.05	10.05	.30	.33	.37	.43	.39	.34	.40	.53	.24	.36	.30	.38	.34	—		
15. Attention <sup>n</sup>	37.02	12.86	0.00	1.00	.38	.39	.47	.50	.50	.44	.47	.45	.24	.41	.36	.33	.33	.51	—	
16. Processing speed <sup>o</sup>	30.74	5.63	98.27	15.48	.43	.37	.47	.32	.32	.30	.32	.24	.32	.30	.29	.23	.25	.26	.41	—

Note. PS = problem solving.

<sup>a</sup> Addition Fact Fluency. <sup>b</sup> Subtraction Fact Fluency. <sup>c</sup> Test of Computational Fluency. <sup>d</sup> Simple Word Problems. <sup>e</sup> Algorithmic Word Problems. <sup>f</sup> Complex Word Problems. <sup>g</sup> A factor score across the Woodcock Diagnostic Reading Battery Listening Comprehension, Test of Language Development-Primary Grammatic Closure, and Wechsler Abbreviated Scale of Intelligence (WASI) Vocabulary. <sup>h</sup> Woodcock-Johnson III Tests of Achievement (WJ-III) Retrieval Fluency (W score). <sup>i</sup> WJ-III Concept Formation. <sup>j</sup> WASI Matrix Reasoning. <sup>k</sup> Working Memory Test Battery for Children Listening Recall. <sup>l</sup> WJ-III Numbers Reversed. <sup>m</sup> Woodcock Reading Mastery Tests-Revised Word Identification. <sup>n</sup> Strengths and Weaknesses of Attention-Deficit/Hyperactivity Disorder-Symptoms and Normal-Behavior Rating Scale. <sup>o</sup> WJ-III Visual Matching.

Table 2  
Full and Pruned Regression Models Predicting Computation<sup>a</sup> Performance ( $n = 919$ )

Predictor	Full			Pruned		
	<i>B</i>	<i>SE</i>	<i>t</i>	<i>B</i>	<i>SE</i>	<i>t</i>
Intercept	-0.03	0.05	0.62	-0.01	0.05	0.30
Problem solving <sup>b</sup>	0.45	0.04	11.18***	0.44	0.04	11.41***
Language <sup>c</sup>	-0.08	0.04	-2.13*	-0.08	0.03	-2.37*
Semantic retrieval <sup>d</sup>	0.03	0.03	0.99			
Concept formation <sup>e</sup>	-0.01	0.03	-0.33			
Matrix reasoning <sup>f</sup>	-0.03	0.03	-0.88			
Verbal working memory <sup>g</sup>	-0.04	0.03	-1.24			
Numerical working memory <sup>h</sup>	-0.01	0.03	-0.17	-0.02	0.03	-0.62
Word ID <sup>i</sup>	0.11	0.03	3.15*	0.10	0.03	3.00***
Attention <sup>j</sup>	0.12	0.03	3.50***	0.11	0.03	3.39***
Processing speed <sup>k</sup>	0.25	0.03	8.22***	0.26	0.03	8.90***
Problem solving-Q	0.04	0.02	1.87	-0.04	0.02	-1.96
Language-Q	-0.02	0.02	-0.87			
Semantic retrieval-Q	0.01	0.02	0.55			
Matrix reasoning-Q	0.00	0.03	0.07			
Numerical working memory-Q	0.05	0.02	2.77*	0.05	0.02	2.95*
Processing speed-Q	0.03	0.02	1.63	0.03	0.02	1.63

Note. Full,  $F(16, 902) = 42.22, p < .001, R^2 = .43$ . Pruned,  $F(8, 910) = 83.54, p < .001, R^2 = .42$ . Q indicates the quadratic term.

<sup>a</sup> Addition Fact Fluency, Subtraction Fact Fluency, Test of Computational Fluency. <sup>b</sup> Simple Word Problems, Algorithmic Word Problems, and Complex Word Problems. <sup>c</sup> Woodcock Diagnostic Reading Battery Listening Comprehension, Test of Language Development-Primary Grammatical Closure, and Wechsler Abbreviated Scale of Intelligence (WASI) Vocabulary. <sup>d</sup> Woodcock-Johnson III Tests of Achievement (WJ-III) Retrieval Fluency (W score). <sup>e</sup> WJ-III Concept Formation. <sup>f</sup> WASI Matrix Reasoning. <sup>g</sup> Working Memory Test Battery for Children Listening Recall. <sup>h</sup> WJ-III Numbers Reversed. <sup>i</sup> Woodcock Reading Mastery Tests-Revised Word Identification. <sup>j</sup> Strengths and Weaknesses of Attention-Deficit/Hyperactivity Disorder-Symptoms and Normal-Behavior Rating Scale. <sup>k</sup> WJ-III Visual Matching.

\*  $p < .05$ . \*\*\*  $p < .001$ .

resulted in 415 ND, 35 CD, 33 PD, and 64 CPD. All remaining analyses were conducted on this subset of 547 students classified as ND, CD, PD, or CPD (i.e., not in the buffer zone). Because academic performance occurs on a continuum, cutoffs for denoting difficulty or lack thereof are necessarily arbitrary, as is the designation of learning disability in the schools. The 40th percentile is used commonly in the research literature as a cut-point for designating lack of difficulty. Cutoffs for designating difficulty or disability vary more in the literature. We selected the 15th percentile because it is useful for understanding disability as practiced in the schools. In Table 4, we show standard score means and standard deviations for the nationally normed math variables we had available and for the cognitive dimensions on which we had nationally normed data.

### Sociodemographic Comparisons

In Table 5, we show frequencies for gender, subsidized lunch, and ethnicity by difficulty status. For each variable, we partitioned the contingency tables to run a series of chi-square tests for determining (a) whether ND differed from students with difficulty (ND vs. CD/PD/CPD), (b) whether CD differed from either variant of problem-solving difficulty (CD vs. PD/CPD), and (c) whether the two variants of problem-solving difficulty differed (PD vs. CPD). To control for multiple comparisons, we adjusted the value of alpha by the number of contrasts on each index (i.e., tested at  $p = .05 / 3 = .0167$ ). There was no significant relation between

gender and math difficulty status:  $\chi^2(1, N = 547) = 0.02, ns$ ;  $\chi^2(1, N = 132) = 0.01, ns$ ; and  $\chi^2(1, N = 97) = 0.50, ns$ , for the three contrasts, respectively. For the proportion of subsidized lunch, however, an interesting set of relations emerged. As might be anticipated, students without math difficulty were significantly less likely to receive subsidized lunch than were students with math difficulty,  $\chi^2(1, N = 547) = 6.82, p = .009$ . More interesting is the fact that students with computational difficulty were significantly less likely to receive subsidized lunch than were students with problem-solving difficulty (PD or CPD),  $\chi^2(1, N = 132) = 7.01, p = .008$ , even though students with PD alone and those with CPD were comparably likely to receive subsidized lunch,  $\chi^2(1, N = 97) = 0.46, p = .50$ . The same pattern emerged for ethnicity. The distribution of ethnicity differed between students with and without difficulty,  $\chi^2(5, N = 547) = 33.13, p < .001$ . The distribution of ethnicity also differed for students with CD versus those with problem-solving difficulty (PD or CPD),  $\chi^2(5, N = 132) = 22.89, p < .001$ , even though students with the two variants of problem-solving difficulty (PD vs. CPD) were similarly distributed across the ethnicity categories,  $\chi^2(5, N = 97) = 7.71, p = .17$ .

### Group Comparisons on Cognitive Dimensions

Table 6 presents z-score means and standard deviations, by difficulty status, on the computation and problem-solving factor scores and on language (factor score), semantic retrieval fluency, concept formation, matrix reasoning, verbal working memory,



Table 3  
Full and Pruned Regression Models Predicting Problem-Solving<sup>a</sup> Performance ( $n = 919$ )

Predictor	Full			Pruned		
	<i>B</i>	<i>SE</i>	<i>t</i>	<i>B</i>	<i>SE</i>	<i>t</i>
Intercept	-0.05	0.04	1.25	-0.07	0.03	2.32*
Computation <sup>b</sup>	0.31	0.03	11.58***	0.29	0.03	11.77***
Language <sup>c</sup>	0.23	0.03	7.29***	0.24	0.03	9.05***
Semantic retrieval <sup>d</sup>	-0.01	0.03	-0.25			
Concept formation <sup>e</sup>	0.16	0.03	5.84***	0.16	0.03	5.95***
Matrix reasoning <sup>f</sup>	0.13	0.03	5.04***	0.14	0.02	5.44***
Verbal working memory <sup>g</sup>	0.03	0.03	1.21			
Numerical working memory <sup>h</sup>	0.05	0.03	1.92	0.05	0.02	2.14*
Word ID <sup>i</sup>	0.02	0.03	0.71			
Attention <sup>j</sup>	0.18	0.03	6.22***	0.17	0.03	6.36***
Processing speed <sup>k</sup>	-0.03	0.03	-1.26			
Language-Q	0.03	0.02	1.90	0.03	0.02	2.00*
Concept formation-Q	0.01	0.02	0.90			
Matrix reasoning-Q	-0.02	0.02	-0.72			
Numerical working memory-Q	0.03	0.02	1.89			
Attention-Q	0.04	0.02	2.32*	0.04	0.02	2.18*

Note. Full,  $F(15, 903) = 81.96$ ,  $p < .001$ ,  $R^2 = .58$ . Pruned,  $F(8, 910) = 152.55$ ,  $p < .001$ ,  $R^2 = .57$ . Q indicates the quadratic term.

<sup>a</sup> Simple Word Problems, Algorithmic Word Problems, and Complex Word Problems. <sup>b</sup> Addition Fact Fluency, Subtraction Fact Fluency, Test of Computational Fluency. <sup>c</sup> Woodcock Diagnostic Reading Battery Listening Comprehension, Test of Language Development-Primary Grammatical Closure, and Wechsler Abbreviated Scale of Intelligence (WASI) Vocabulary. <sup>d</sup> Woodcock-Johnson III Tests of Achievement (WJ-III) Retrieval Fluency (W score). <sup>e</sup> WJ-III Concept Formation. <sup>f</sup> WASI Matrix Reasoning. <sup>g</sup> Working Memory Test Battery for Children Listening Recall. <sup>h</sup> WJ-III Numbers Reversed. <sup>i</sup> Woodcock Reading Mastery Tests-Revised Word Identification. <sup>j</sup> Strengths and Weaknesses of Attention-Deficit/Hyperactivity Disorder-Symptoms and Normal-Behavior Rating Scale. <sup>k</sup> WJ-III Visual Matching.

\*  $p < .05$ . \*\*\*  $p < .001$ .

numerical working memory, word identification, attentive behavior (factor score), and processing speed.

**Preliminary analysis of clustering effects.** Before choosing an analytic model, we examined the extent of clustering of children in classes and schools. With clustering, the independence assumption of analysis of variance may be violated, possibly leading to spurious significance levels (Raudenbush & Bryk, 2002). Strong clustering would necessitate a multilevel model rather than a repeated measures approach. Variance components were estimated with SAS PROC MIXED (Littell, Milliken, Stroup, Wolfinger, & Schabenger, 2006). The resulting intraclass correlations showed how much of the total variance in the variables of interest (i.e., the nine cognitive dimensions) was explained by the clustering of children in classroom and school. The effect for school explained 0% of the variance, and the effect for classroom (nested in school) explained less than 5%. Raudenbush and Liu's (2000) ad hoc standards deem 5%, 10%, and 15% as small, medium, and large.

**Overall analysis.** Because the intraclass correlations in this database were small to nonexistent, we conducted an initial profile analysis using a two-way analysis of variance. The between-subjects factor was math difficulty status (ND vs. CD vs. PD vs. CPD); the within-subjects factor was cognitive dimension ( $z$  score on language vs. semantic retrieval fluency vs. concept formation vs. matrix reasoning vs. verbal working memory vs. numerical working memory vs. word identification vs. attentive behavior vs. processing speed). The interaction between math difficulty status and cognitive dimension was significant,  $F(24, 4337) = 4.67$ ,  $p < .0001$ .

To help interpret the interaction between math difficulty status and cognitive dimension, we plotted  $z$  scores on the nine cognitive dimensions for each of the four difficulty status groups (see Figure 1). As shown, the ND and CD groups performed at a higher level than did the PD and CPD groups. In addition, the difficulty status by cognitive dimension interaction appeared to be evident in the variations in the profile shape across the nine cognitive dimensions as a function of the difficulty status group.

**Univariate follow-ups to the interaction.** Because the elevation effects in Figure 1 were striking and because many studies have compared univariate differences among math difficulty groups, we initially conducted follow-up tests using Fisher's least significant difference, with math difficulty status as the factor (Seaman, Levin, & Serlin, 1991). Alpha was adjusted for six contrasts per measure, comparing each difficulty status group to all others ( $p = .05 / 6 = .008$ ). Results of these follow-up tests appear under the labels on the horizontal axis in Figure 1. Symbols for the four groups (see key) appear under each cognitive dimension. Groups joined by a horizontal line were not significantly different from one another. The univariate test on each cognitive dimension was significant ( $p < .0001$ ). To help evaluate the magnitude of the univariate differences, we computed effect sizes for each variable (see Table 7) by dividing the difference between group means by the standard deviation pooled across the two groups in the comparison (Hedges & Olkin, 1985). The complicated pattern of differences that emerged is addressed in the Discussion.

**Profile analysis of shape effects.** The interpretation of a pattern like the one shown in Figure 1 is complicated. The univariate

Table 4  
Standard Score Performance by Difficulty Status

Variable	Difficulty status							
	ND ( <i>n</i> = 415)		CD ( <i>n</i> = 35)		PD ( <i>n</i> = 33)		CPD ( <i>n</i> = 64)	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
TerraNova	58.80	14.97	22.06	14.34	24.83	15.16	17.00	15.60
Applied problems <sup>a</sup>	109.65	12.80	100.31	11.15	93.30	7.72	88.17	9.00
Grammatic <sup>b</sup>	89.43	11.61	86.14	10.30	78.48	8.15	78.83	6.59
Listening <sup>c</sup>	101.92	18.37	104.31	14.11	84.88	11.71	84.36	12.45
Vocabulary <sup>d</sup>	50.88	9.89	49.49	10.20	40.33	6.64	40.75	7.77
Semantic retrieval <sup>e</sup>	96.94	12.94	93.26	12.92	92.36	12.94	86.00	14.07
Concept formation <sup>f</sup>	98.74	12.03	91.80	13.58	84.82	10.59	81.48	10.95
Matrix reasoning <sup>g</sup>	53.35	10.27	49.06	9.94	42.15	9.63	42.45	9.68
Verbal working memory <sup>h</sup>	97.72	14.59	90.23	15.47	83.27	16.07	83.66	15.62
Numerical working memory <sup>i</sup>	99.73	14.13	95.20	10.43	90.30	12.19	89.08	10.19
Word ID <sup>j</sup>	104.93	10.50	100.69	9.45	94.94	6.97	93.08	6.73
Processing speed <sup>k</sup>	104.35	14.83	92.37	12.76	96.42	17.71	87.56	13.85

Note. TerraNova is normal curve equivalents on the TerraNova. All other scores are standard scores. Standard scores are *M* = 100, *SD* = 15, except vocabulary and matrix reasoning, for which *M* = 50 and *SD* = 10. ND = no difficulty; CD = computational difficulty; PD = problem-solving difficulty; CPD = computational and problem-solving difficulty.

<sup>a</sup> Woodcock-Johnson III Tests of Achievement (WJ-III) Applied Problems. Applied Problems assesses a variety of math domains with only a limited number of word problems assessed in the third-grade range. <sup>b</sup> Test of Language Development-Primary Grammatic Closure. <sup>c</sup> Woodcock Diagnostic Reading Battery Listening Comprehension. <sup>d</sup> Wechsler Abbreviated Scale of Intelligence (WASI) Vocabulary. <sup>e</sup> WJ-III Retrieval Fluency. <sup>f</sup> WJ-III Concept Formation. <sup>g</sup> WASI Matrix Reasoning. <sup>h</sup> Working Memory Test Battery for Children Listening Recall. <sup>i</sup> WJ-III Numbers Reversed. <sup>j</sup> Woodcock Reading Mastery Tests-Revised Word Identification. <sup>k</sup> WJ-III Visual Matching.

tests do not account for relations among the cognitive dimensions, and they confound level and measure effects. A traditional multivariate analysis of variance determines how a set of measures can be combined into a set of  $K - 1$  univariate composites (discriminant functions), which maximally separates groups (Bernstein, Garbin, & Teng, 1988; Harris, 1975). However, each composite comprises elements that involve differences not only in elevation but also in shape, both of which seem apparent in the profiles shown in Figure 1. By contrast, a multivariate profile analysis accounts for correlations among cognitive dimensions and decomposes the univariate composites into components representing elevation, flatness, and shape. The elevation effects are differences among groups averaged across the dimensions. Flatness effects are

differences in dimensions averaged across groups; they indicate whether the profile measures vary or can be represented as a relatively straight line (because we used *z* scores, we did not anticipate flatness effects). When profiles differ in shape, which is analogous to the Difficulty Status  $\times$  Dimension interaction already documented, differences among groups vary depending on cognitive dimension, in which case the elevation and flatness components are not interesting.

Thus, we used multivariate profile analysis to conduct four planned contrasts to explore how the profiles of the math difficulty groups differed. In these contrasts, we compared the ND group to the specific computational difficulty group and then in turn compared the specific computational difficulty group to each of the

Table 5  
Demographics by Difficulty Status

Demographic variable	Difficulty status							
	ND ( <i>n</i> = 415)		CD ( <i>n</i> = 35)		PD ( <i>n</i> = 33)		CPD ( <i>n</i> = 64)	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Male	201	48.4	17	48.6	14	42.4	32	50.0
Subsidized lunch <sup>a</sup>	185	23.4	14	45.2	18	66.7	45	73.8
African American	129	31.1	10	28.6	23	70.0	43	67.2
European American	201	48.4	22	62.9	6	18.2	14	21.9
Hispanic	36	8.7	3	8.6	1	3.0	6	9.4
Kurdish	14	3.4	0	0.0	1	3.0	0	0.0
Other	35	8.4	0	0.0	2	6.1	1	1.6

Note. ND = no difficulty; CD = computational difficulty; PD = problem-solving difficulty; CPD = computational and problem-solving difficulty.

<sup>a</sup> Some schools declined to provide subsidized lunch data, resulting in sample sizes of 363, 31, 27, and 61 for the four difficulty status groups, respectively.



Table 6  
Performance by Difficulty Status

Variable	Difficulty status							
	ND ( <i>n</i> = 415)		CD ( <i>n</i> = 35)		PD ( <i>n</i> = 33)		CPD ( <i>n</i> = 64)	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Basic facts + <sup>a</sup>	0.62	0.80	-1.21	0.43	0.40	0.61	-1.44	0.53
Basic facts - <sup>b</sup>	0.62	0.95	-0.95	0.31	0.09	0.74	-1.02	0.45
Computation <sup>c</sup>	0.76	0.79	-1.09	0.20	0.14	0.62	-1.27	0.30
Simple PS <sup>d</sup>	0.76	0.49	0.36	0.53	-1.33	0.54	-1.75	0.59
Algorithmic PS <sup>e</sup>	0.68	0.96	0.00	0.57	-0.98	0.22	-1.04	0.22
Complex PS <sup>f</sup>	0.56	1.10	-0.18	0.55	-0.84	0.22	-0.78	0.22
Language <sup>g</sup>	0.40	0.89	0.32	0.79	-0.85	0.82	-0.81	0.82
Semantic retrieval <sup>h</sup>	0.27	0.93	0.01	0.93	-0.05	0.93	-0.51	1.01
Concept formation <sup>i</sup>	0.42	0.88	-0.09	1.00	-0.60	0.78	-0.85	0.80
Matrix reasoning <sup>j</sup>	0.37	0.92	-0.01	0.90	-0.64	0.87	-0.61	0.87
Verbal working memory <sup>k</sup>	0.31	0.91	-0.16	0.97	-0.59	1.00	-0.57	0.97
Numerical working memory <sup>l</sup>	0.27	1.01	-0.05	0.74	-0.40	0.87	-0.49	0.73
Word ID <sup>m</sup>	0.39	1.04	-0.04	0.94	-0.61	0.69	-0.79	0.67
Attention <sup>n</sup>	0.55	0.90	-0.47	0.94	-0.59	0.72	-1.07	0.53
Processing speed <sup>o</sup>	0.39	0.96	-0.38	0.82	-0.12	1.14	-0.69	0.89

Note. Performance is expressed as *z* scores in relation to the representative sample of 919. ND = no difficulty; CD = computational difficulty; PD = problem-solving difficulty; CPD = computational and problem-solving difficulty.

<sup>a</sup> Addition Fact Fluency. <sup>b</sup> Subtraction Fact Fluency. <sup>c</sup> Test of Computational Fluency. <sup>d</sup> Simple Word Problems. <sup>e</sup> Algorithmic Word Problems. <sup>f</sup> Complex Word Problems. <sup>g</sup> A factor score across the Woodcock Diagnostic Reading Battery Listening Comprehension, Test of Language Development-Primary Grammatical Closure, and Wechsler Abbreviated Scale of Intelligence (WASI) Vocabulary. <sup>h</sup> Woodcock-Johnson III Tests of Achievement (WJ-III) Retrieval Fluency. <sup>i</sup> WJ-III Concept Formation. <sup>j</sup> WASI Matrix Reasoning. <sup>k</sup> Working Memory Test Battery for Children Listening Recall. <sup>l</sup> WJ-III Numbers Reversed. <sup>m</sup> Woodcock Reading Mastery Tests—Revised Word Identification. <sup>n</sup> Strengths and Weaknesses of Attention-Deficit/Hyperactivity Disorder-Symptoms and Normal-Behavior Rating Scale. <sup>o</sup> WJ-III Visual Matching.

groups involving problem-solving difficulty, thereby reducing the need to compare the ND and problem-solving difficulty groups directly (especially given that the ND group performed at a much higher level than the CPD and PD groups, as shown in Figure 1). We also compared the specific problem-solving difficulty group to the group that manifested both forms of difficulty. To control for Type 1 error (given four planned contrasts), we adjusted the critical value of alpha to .0125 (.05 / 4); however, because the power for each comparison differed, effect sizes (eta-squared) were also considered.

Although separate effects for the elevation, flatness, and shape effects might be explored in a profile analysis approach, the large interaction already documented in the univariate analysis between difficulty status and cognitive dimension mitigated against pursuing elevation or flatness effects. Table 8 presents the interaction contrasts exploring the shape effects. As shown, only the contrasts for ND versus CD and for CD versus PD met the critical alpha level. Although the contrasts between PD and CPD and between CD and CPD did not meet the critical alpha level, the effect sizes (both ~0.15) were slightly larger than for the ND versus CD contrast (0.08), the latter having more power. Effect sizes were in the small to medium range.

In interpreting the shape effects, we note that the overall tests for elevation, flatness, and shape are not dependent on the ordering of the dimensions. However, interpretation of individual dimensions may depend on the ordering and the location of pairs of dimen-

sions in the profile. To determine how profile dimensions contribute to the interaction regardless of the ordering of the dimensions, we followed a commonly recommended procedure for the interpretation of profile analysis multivariate analysis of variance: inspection of the canonical structure matrix (Harris, 1975). In following up the shape effect, it is also necessary to remove the effect of elevation (Bernstein et al., 1988; Fletcher et al., 1994). This was done by computing the residuals in a model, which included only the main effects of group and cognitive dimension; the result reduced the elevation of each group to approximately zero. Any variation among group means on the cognitive variables is then due only to the effect of shape.

Figure 2 shows the elevation-adjusted shape profile for each of the four groups along the nine cognitive dimensions. Note that the relation of the groups on the *y*-axis differs substantially from the relations manifested in Figure 1. This is because the elevation differences that dominated Figure 1 have been removed in Figure 2. In Table 9, we present the canonical structure matrix for each of the four planned contrasts. Within the canonical structure matrix, simple correlations are computed for each variable with the discriminant function representing the multivariate effect for shape, adjusted for elevation (Huberty, 1975). The positive or negative value of the correlations reflects the pattern of mean differences between the contrasted groups. For interpretive purposes, we consider both the magnitude and the direction of these correlations as well as the significance of the univariate test (*F*

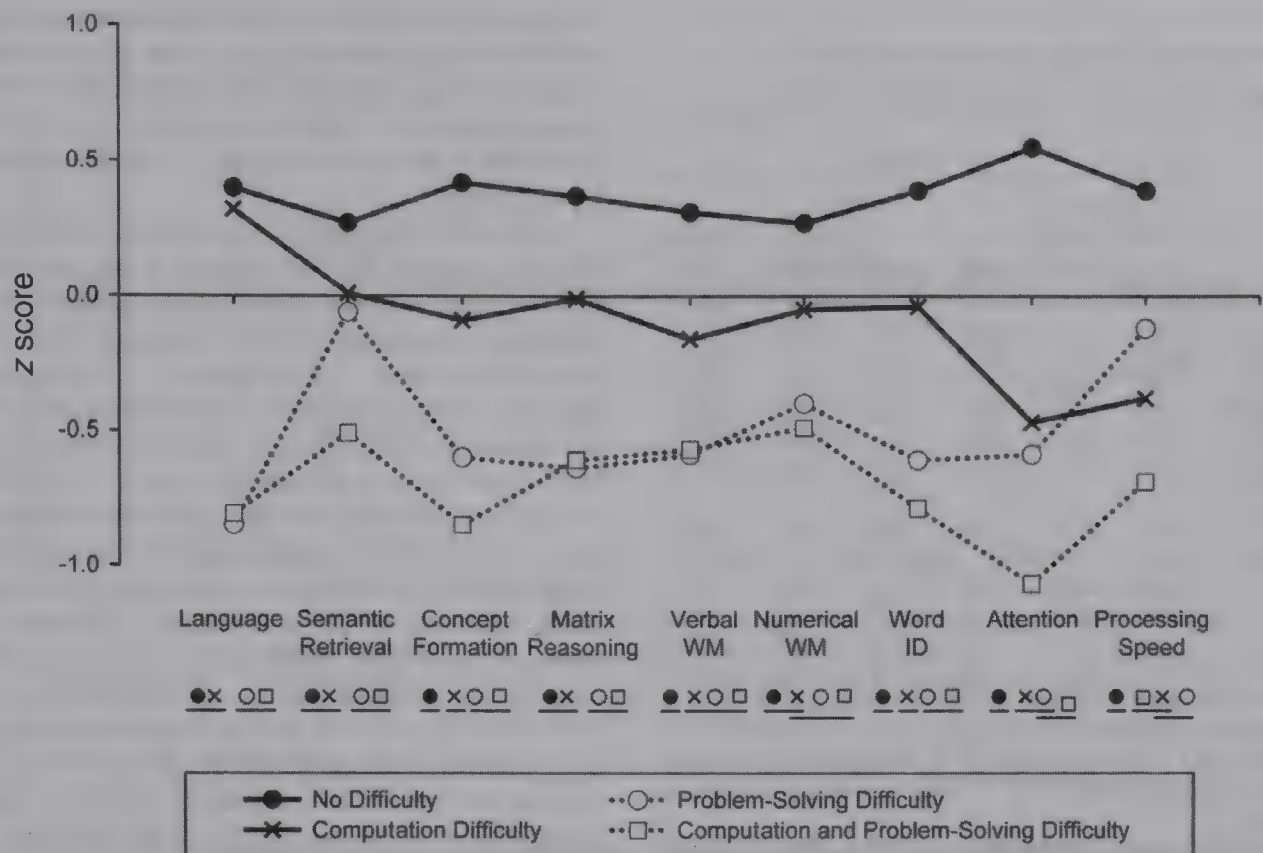


Figure 1. *z* scores on nine cognitive dimensions by difficulty status. WM = working memory; ID = identification.

value) associated with each cognitive variable for each pair of contrasts (see asterisks in Table 9, where the critical *p* value has been adjusted to .006 to account for nine univariate dimensions).

As reflected in the canonical structure correlations shown in Table 9, the contrast between the ND and CD groups was accounted for by language (−.44), attentive behavior (.60), and processing speed (.32), which were more heavily weighted than other variables. Accordingly, in Figure 2, the difference in language (positive direction for CD relative to other dimensions), attentive behavior (negative direction for CD relative to other dimensions), and processing speed (negative direction for CD relative to other dimensions) most clearly differentiates the shape of the ND and CD profiles. Among language, attentive behavior, and processing speed, attentive behavior was the most reliable correlate of the shape effect across different methods of interpreting the contributions of the dimensions.

As might be expected, therefore, for the comparison between the PD and CPD groups, Table 9 shows the highest canonical structure coefficients for attentive behavior (.54) and processing speed (.44). In Figure 2, these differences in attentive behavior (negative direction for CPD relative to most other dimensions) and processing speed (positive direction for PD relative to most other dimensions) are striking aspects of the profile and also clearly contrast with the profile for the CPD group, in which attentive behavior was a negative dimension and processing speed is neutral. Again, attentive behavior was the most reliable correlate of the shape effect across different methods of interpreting the contributions of the dimensions.

By contrast, for the comparison between PD and CD, language (−.70) and processing speed (.51) were the variables contributing

to the shape effect and reflect the negative direction of language relative to other dimensions for the PD group and the positive direction in language relative to other dimensions for the CD group. In contrast, processing speed was a negative dimension for the CD group but positive for the PD group. Figure 2 shows that semantic retrieval fluency was also positive in the PD group, a likely suppressor effect given the large standardized coefficient for semantic retrieval fluency (.74). Language was the most reliable correlate of the shape effect across different methods of interpreting the contributions of the dimensions. Finally, and in keeping with the contrast between PD and CD, the cognitive dimension accounting for the contrast between the CPD and CD groups was language (−.72). The standardized coefficients showed a similar pattern. Again, this variable moved in opposite directions in the profiles for the two groups (see Figure 2).

## Discussion

The purposes of this study were to explore the overlap between difficulty with two aspects of primary-grade mathematical cognition—computation and problem solving—and to examine how demographic and cognitive profiles differ among subgroups with difficulty in one, the other, both, or neither. The goal was to gain insight into whether these domains are shared or distinct aspects of mathematical cognition in extreme groups. This issue is not only theoretically important but also has implications in terms of identifying math disability, as specified in the 2004 reauthorization of the Individuals with Disabilities Education Act, and for designing effective methods for preventing and remediating math difficulty.



Table 7  
Effect Sizes (in Absolute Values) for Math Variables and Cognitive Dimensions as a Function of Difficulty Status

Variable	Contrast					
	ND vs.			CD vs.		PD vs. CPD
	CD	PD	CPD	PD	CPD	
Computation <sup>a</sup>	2.80	0.73	3.16	4.83	0.86	5.13
Problem solving <sup>b</sup>	0.89	2.48	2.78	4.13	5.27	0.68
Language <sup>c</sup>	0.09	1.45	1.37	1.44	1.40	0.05
Semantic retrieval <sup>d</sup>	0.28	0.34	0.83	0.06	0.53	0.46
Concept formation <sup>e</sup>	0.57	1.17	1.45	0.60	0.87	0.32
Matrix reasoning <sup>f</sup>	0.39	1.09	1.08	0.70	0.67	0.03
Verbal working memory <sup>g</sup>	0.52	0.98	0.96	0.44	0.42	0.02
Numerical working memory <sup>h</sup>	0.32	0.67	0.78	0.43	0.59	0.12
Word ID <sup>i</sup>	0.42	0.98	1.19	0.70	0.97	0.26
Attention <sup>j</sup>	1.13	1.28	1.64	0.15	0.69	0.80
Processing speed <sup>k</sup>	0.75	0.68	1.46	0.16	0.52	0.58

Note. Performance is expressed as *z* scores in relation to the representative sample of 919. ND = no difficulty; CD = computational difficulty; PD = problem-solving difficulty; CPD = computational and problem-solving difficulty.

<sup>a</sup> A factor score (Addition Fact Fluency, Subtraction Fact Fluency, Test of Computational Fluency). <sup>b</sup> A factor score (Simple Word Problems, Algorithmic Word Problems, Complex Word Problems). <sup>c</sup> A factor score across the Woodcock Diagnostic Reading Battery Listening Comprehension, Test of Language Development–Primary Grammatical Closure, and Wechsler Abbreviated Scale of Intelligence (WASI) Vocabulary. <sup>d</sup> Woodcock–Johnson III Tests of Achievement (WJ–III) Retrieval Fluency. <sup>e</sup> WJ–III Concept Formation. <sup>f</sup> WASI Matrix Reasoning. <sup>g</sup> Working Memory Test Battery for Children Listening Recall. <sup>h</sup> WJ–III Numbers Reversed. <sup>i</sup> Woodcock Reading Mastery Tests—Revised Word Identification. <sup>j</sup> Strengths and Weaknesses of Attention-Deficit/Hyperactivity Disorder-Symptoms and Normal-Behavior Rating Scale. <sup>k</sup> WJ–III Visual Matching.

With respect to overlap, results revealed that difficulty in one domain did not necessarily align with difficulty in the other. This is understandable because correlations between computational and problem-solving skill, as demonstrated elsewhere (e.g., Fuchs et al., 2006; Swanson & Beebe-Frankenberger, 2004), were only moderate, ranging from .30 to .49. In fact, difficulty occurred in a single math domain as frequently as across math domains. Moreover, specific difficulty was distributed across the two domains with almost identical prevalence.

In a related way, the demographic profiles of the groups also suggest that performance in these two domains of mathematical cognition may be distinct. The demographic profiles of students with specific computational difficulty, in terms of poverty and ethnic background, were more similar to those of students without difficulty than to those of students with problem-solving difficulty or with concurrent difficulty. This suggests that the contextual variables associated with poverty or race exert little effect over the development of computational difficulty. By contrast, students with problem-solving difficulty were significantly poorer and disproportionately more likely to be African American compared to students with specific computational difficulty. This was true regardless of whether problem-solving deficits occurred alone or in combination with computational deficits, with no significant

demographic differences between students experiencing specific problem-solving difficulty and those experiencing concurrent problem-solving and computational difficulty. This finding indicates that early or ongoing experience outside of school may account for variance in building a strong foundation for mathematical problem solving.

To identify what kinds of experience outside of school may be key, it is useful to consider analysis of the cognitive dimensions. The regression analyses indicate that a different set of cognitive predictors is associated with the math outcomes, suggesting that these two domains of mathematical cognition may be distinct. However, because the focus of the present study was to extend understanding about difficulty with computation versus problem solving, and because the regressions only inform about the effect of each variable when the other variables are controlled at their means, it is important to examine performance specifically at the lower ranges of performance on the math outcomes.

The univariate profile analyses, which addressed mean differences among difficulty status groups on each cognitive dimension, indicate that language and word identification clearly distinguished problem-solving from computational difficulty. Students with problem-solving difficulty, regardless of whether the problem-solving difficulty occurred alone or in combination with computational difficulty, scored reliably lower than students with neither form of difficulty and lower than those who experienced computational deficits alone. Moreover, students with computational difficulty were statistically indistinguishable from students without difficulty, and students with problem-solving difficulty alone were statistically comparable to students with problem-solving difficulty that occurred in combination with computational difficulty. Two additional variables, concept formation and matrix reasoning, also served to distinguish problem-solving difficulty from computational difficulty (with PD and CPD comparable to each other, and both lower than students with CD and ND, although on these dimensions, students with specific computational difficulty were reliably lower performing than students without difficulty).

It is therefore interesting to consider these abilities in light of the major distinction between mathematical computation and problem solving: the addition of linguistic information that requires individuals to construct a problem model. That is, whereas a computation problem is already set up for solution, a word problem requires students to use text to identify missing information, construct the number sentence, and derive the calculation problem for finding the missing information. This transparent difference would

Table 8  
Math Difficulty Status × Cognitive Dimension Interactions for Four Planned Comparison

Contrast	<i>F</i>	<i>dfs</i>	<i>p</i>	$\eta^2$
ND vs. CD	4.22	8, 441	.0001	.08
PD vs. CPD	1.85	8, 87	.08	.15
CD vs. PD	4.75	8, 59	.0002	.39
CD vs. CPD	2.26	8, 89	.03	.16

Note. ND = no difficulty; CD = computational difficulty; PD = problem-solving difficulty; CPD = computational and problem-solving difficulty.

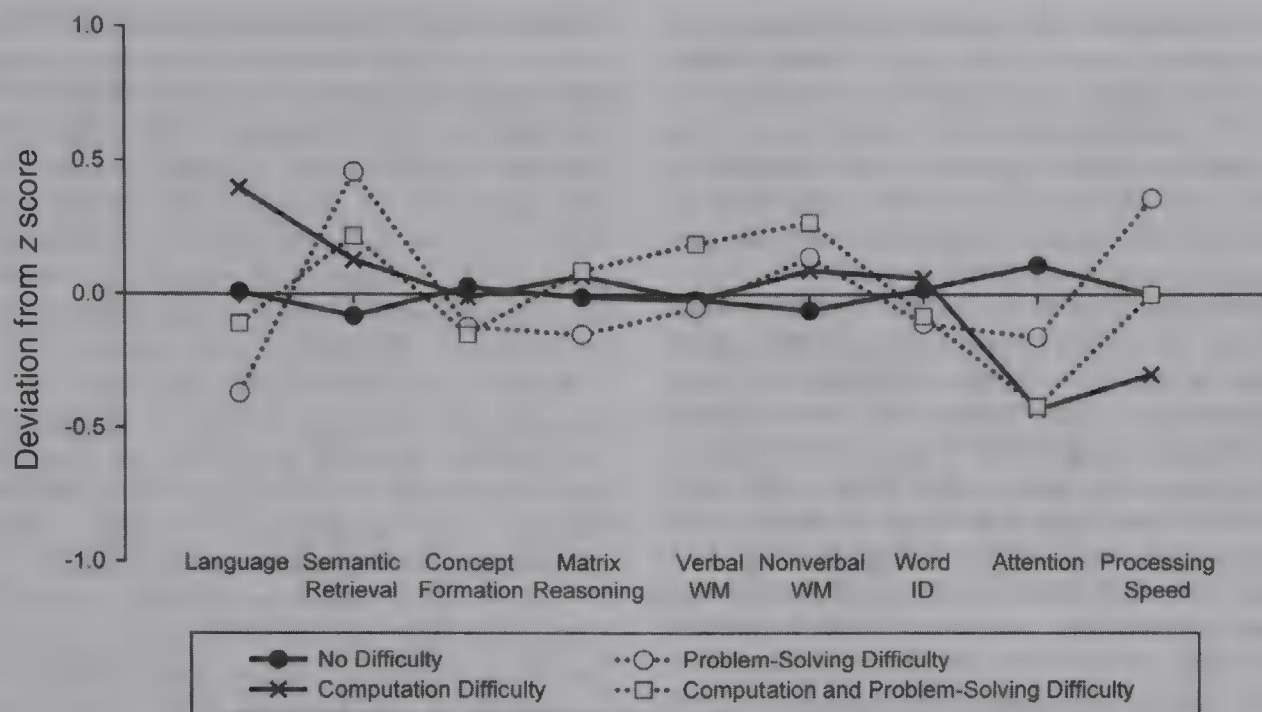


Figure 2. Shape effects by difficulty status. WM = working memory; ID = identification.

seem to alter the nature of the task, and findings corroborate such a hypothesis.

With respect to the contribution of reading skill, although word problems were read aloud to students, with repeated opportunities for rereading whenever students requested, students had the written problem available until they completed it; so, we note that independent, skilled reading may support continuous access to text. In a different way, because the development of word identification skill is facilitated by vocabulary knowledge (cf. Perfetti, 1992), the link between word identification and problem-solving skill suggests that language may play a role in math problem solving. This finding was in fact documented in the univariate analyses. It stands to reason

that the ability to make sense of language, as reflected indirectly by word recognition skill and as reflected directly by our language factor score (i.e., grammatic closure, listening comprehension, and vocabulary), should help students cope with narratives in the service of building problem models, and findings corroborate previous work about the role language plays in problem-solving skill (e.g., Fuchs et al., 2006). It also makes sense that nonverbal problem-solving skill, as reflected in concept formation and matrix reasoning, should underlie mathematical problem-solving skill. This corroborates previous work (Fuchs et al., 2005, 2006) and is interpretable because mathematical problem solving requires students not only to build a problem model but also to distinguish relevant from irrelevant

Table 9

Canonical Structure Correlations for Each Cognitive Dimension for the Shape Effect

Variable	ND vs. CD	PD vs. CPD	PD vs. CD	CPD vs. CD
Language <sup>a</sup>	-.44	-.38	-.70*	-.72*
Semantic retrieval <sup>b</sup>	.22	.32	.23	.07
Concept formation <sup>c</sup>	.04	.06	-.08	-.19
Matrix reasoning <sup>d</sup>	-.09	-.30	-.16	.01
Verbal working memory <sup>e</sup>	-.01	-.32	.22	.26
Numerical working memory <sup>f</sup>	-.15	-.19	.04	.26
Word ID <sup>g</sup>	-.05	-.04	-.13	-.20
Attention <sup>h</sup>	.60*	.54*	.21	.01
Processing speed <sup>i</sup>	.32	.44	.51	.38

Note. ND = no difficulty; CD = computational difficulty; PD = problem-solving difficulty; CPD = computational and problem-solving difficulty.

<sup>a</sup> Woodcock Diagnostic Reading Battery Listening Comprehension, Test of Language Development-Primary Grammatical Closure, and Wechsler Abbreviated Scale of Intelligence (WASI) Vocabulary. <sup>b</sup> Woodcock-Johnson III Tests of Achievement (WJ-III) Retrieval Fluency (W score). <sup>c</sup> WJ-III Concept Formation. <sup>d</sup> WASI Matrix Reasoning. <sup>e</sup> Working Memory Test Battery for Children Listening Recall. <sup>f</sup> WJ-III Numbers Reversed. <sup>g</sup> Woodcock Reading Mastery Tests-Revised Word Identification. <sup>h</sup> Strengths and Weaknesses of Attention-Deficit/Hyperactivity Disorder-Symptoms and Normal-Behavior Rating Scale. <sup>i</sup> WJ-III Visual Matching.

\*  $p < .006$ .



information and to determine how numerical quantities fit into the slots of the problem model (cf. Kintsch & Greeno, 1985).

For the remaining cognitive dimensions, the univariate results were less clear. On verbal and numerical working memory, the performance of the three difficulty groups was indistinguishable. However, on numerical working memory (but not on verbal working memory), students with specific computational difficulty performed comparably to their peers without difficulty.

In terms of working memory, or the capacity to maintain target memory items while processing an additional task (Daneman & Carpenter, 1980), a body of work has established links with computation (Fuchs et al., 2005; Geary et al., 1991; Hitch & McAuley, 1991; Siegel & Linder, 1984; Webster, 1979; Wilson & Swanson, 2001) and problem solving (e.g., Fuchs et al., 2005; LeBlanc & Weber-Russell, 1996; Passolunghi & Siegel, 2004; Swanson & Beebe-Frankenberger, 2004; Swanson & Sachse-Lee, 2001). Although other studies have raised questions about the tenability of this association (e.g., Fuchs et al., 2006; Swanson et al., 1993), the univariate results of the present study corroborate a role for working memory, verbal as well as numerical, in both computational and problem-solving difficulty. For computation, students must hold terms and operators in working memory while using various counting strategies to arrive at the correct answer. Over time, repeated associations for the problem stem and its answer, achieved via successful counting, result in representations in long-term memory, which in turn facilitate fluent procedural computation. With respect to word problems, as described by Kintsch and Greeno (1985), when processing a problem narrative, students formulate new sets to construct a problem model. When a proposition that triggers a set-building strategy is completed, the appropriate set is formed and the relevant propositions are assigned their places in the various slots of the set schema. As new sets are formed, previous sets that had been active in the memory buffer are displaced, illustrating the potential importance of working memory. Our univariate analyses lend empirical support for the contribution of working memory to both aspects of mathematical cognition: computation and problem solving. It is important to note, however, that the multivariate profile analyses indicate that working memory is primarily related to elevation (severity), not to shape.

In any case, results of the univariate analyses suggest a pattern of more pervasive cognitive involvement for problem-solving difficulty. Yet, because univariate analyses fail to account for relations among the cognitive dimensions and because in the univariate analyses, elevation and shape are confounded, it is important to consider results of the multivariate profile analyses. At a general level, by explicating the patterns of cognitive abilities associated with computational and problem-solving difficulty, the shape analyses highlight two important notions: (a) that computational difficulty may be distinct from problem-solving difficulty and (b) that the cognitive dimensions associated with performance in a single math domain may also be associated with both computation and problem solving when difficulties occur concurrently. At the same time, the multivariate profile results, which isolate the effects of shape from the effects of elevation, serve to pinpoint more specifically and narrowly which cognitive differences between groups matter. In this way, results demonstrate the need to rely on multivariate, rather than univariate, approaches in the study of mathematics.

Within the multivariate interpretations of the profile analyses, three cognitive dimensions emerged as central to the distinction between computational and problem-solving difficulty. The dominant role of language deficits was substantiated for problem-solving difficulty. Language was the cognitive dimension that served to distinguish the specific problem-solving difficulty group from the specific computational difficulty group and to distinguish the group with concurrent difficulty across problem solving and computation again from the specific computational difficulty group. By contrast, the dominant roles of attentive behavior and processing speed were revealed for computational difficulty, serving to distinguish the specific computational difficulty group from the group without either form of difficulty and to distinguish the group with concurrent difficulty across computation and problem solving from the group with specific problem-solving difficulty.

In the preceding discussion about the univariate analyses, we already considered the role of language in problem solving, but what about the role of attentive behavior and processing speed in computation, which became evident only in the multivariate analyses? Within the univariate analyses, a step-down pattern occurred for attentive behavior, whereby students without difficulty were rated more favorably than the other three groups; students with specific computational difficulty were rated similarly as students with specific problem-solving difficulty but more attentive than students with concurrent difficulty; and students with specific problem-solving difficulty were deemed similarly attentive as students with concurrent difficulty. This ordering, which suggests greater cognitive involvement for problem-solving deficits, is based entirely on elevation differences among groups. By contrast, the multivariate analyses specifically consider the shape (i.e., the profile) by removing the elevation effects. These multivariate profile analyses demonstrate that attentive behavior is implicated in computational but not problem-solving difficulty. Few studies have considered attentive behavior as a predictor of computational skill, but some previous work has suggested its role (e.g., Fuchs et al., 2005, 2006). Moreover, Swanson (2006) recently substantiated the role of inhibitory control, a form of attention, in the development of computational but not problem-solving skill.

At least two explanations seem possible for the role of attentive behavior in computational difficulty. First, attentive behavior may create the opportunity to persevere with the serial execution required for computational math (Luria, 1980) and thereby enhance performance and improve students' responsiveness to instruction. Alternatively, it is possible that teacher ratings of attentive behavior are clouded by students' academic performance and therefore serve as a proxy for achievement rather than indexing attention. Cirino, Ewing-Cobbs, Barnes, Fuchs, and Fletcher (2007) found that although attentive behavior accounts for unique variance in mathematics performance, removing variance due to behavioral ratings of attention does not alter the relations of cognitive measures and mathematical cognition. Present findings, in which attentive behavior was implicated in specific computational difficulty but not specific problem-solving difficulty, reduce the plausibility that teacher ratings simply serve as a proxy for academic achievement and instead provide the basis for hypothesizing that attentive behavior plays a role in computation and for explor-



ing the underlying nature of the relation with alternative measures of attention.

In terms of processing speed, previous work has suggested that processing speed underlies fluency with math facts. For example, Bull and Johnston (1997) found that processing speed subsumed all of the variance in 7-year-olds' arithmetic skill while controlling for word reading ability, item identification, and short-term memory. Processing speed may facilitate counting speed so that as young children gain speed in counting sets to figure sums and differences, problems are successfully paired with their answers in working memory before decay sets in, such that associations in long-term memory are established (e.g., Geary et al., 1991; Le-maire & Siegler, 1995). In addition, Fuchs et al. (2006) demonstrated that processing speed accounted for unique variance in simple arithmetic but not in mathematical problem solving once the relation between processing speed and arithmetic had been accounted for.

In sum, findings lend support for the hypothesis that computation and problem solving may represent distinct domains of mathematical cognition within students at the lower ranges of performance as might be identified with mathematics learning disabilities in the schools. This lends empirical support for the distinction between computational and mathematics problem-solving learning disabilities specified in the 2004 reauthorization of the Individuals with Disabilities Education Act. Results also suggest that poverty and language play critical roles in the development of problem-solving difficulty and that inattentive behavior and poor processing speed may inhibit the development of computational skill. In addition, despite the more substantial math deficits evidenced for students with concurrent difficulty (i.e., an effect size of 0.86 favoring the computational skill of CD over CPD and an effect size of 0.68 favoring the problem-solving skill of PD over CPD), the cognitive deficiencies associated with math performance in a single domain are also apparent when difficulties occur in both domains: for computation, attention and processing speed (as revealed for CD and for CPD); for problem solving, language (as revealed for PD and for CPD). Similarly, poverty or race is associated with problem-solving difficulty, whether it occurs alone or in combination with computational difficulty, and therefore corroborates the relation between language and these sociodemographic variables. Together, these findings suggest that concurrent difficulty with computation and problem solving may not be a unique form of math disability but represents a comorbid association of difficulties in both domains. Additional research should continue to investigate these issues as well as explore the possible role of other cognitive dimensions, including reading comprehension. Matching groups of students on their areas of math strength may also be a productive line of related work. Further work is also needed using larger samples to yield difficulty status groups with greater numbers of students and using more restrictive cutoffs for denoting difficulty that correspond even more closely to the criteria employed to designate children in schools as having mathematics learning disabilities. In addition, related work using different strategies for measuring computational skill, problem-solving skill, and the nine cognitive dimensions is required to corroborate effects. Furthermore and more generally, present study findings indicate that multivariate analytic approaches are required to untangle the

role of cognitive abilities underlying the development of mathematical skill.

In the meantime, however, we caution practitioners about the potential need to consider computational skill and problem-solving skill separately in diagnosing and instructing students with learning disabilities. We also note that models of mathematical competence should expand focus on mathematical problem solving and explicitly consider the abilities that underlie the development of this form of mathematical competence. Moreover, findings support the importance of studying instructional procedures that may enhance performance specifically in the area of mathematical problem solving, separate from the issue of how to promote computational skill. Finally, the critical importance of assessing computational and problem-solving skills separately for the presence of math difficulties is apparent. Many mathematics assessments are generic and do not adequately attend to the differentiation of these dimensions of mathematical cognition.

## References

- Agness, P. J., & McLone, D. G. (1987). Learning disabilities: A specific look at children with spina bifida. *Insights*, 8–9.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Bernstein, I. H., Garbin, C. P., & Teng, G. K. (1988). *Applied multivariate analyses*. New York: Springer-Verlag.
- Bull, R., & Johnston, R. S. (1997). Children's arithmetical difficulties: Contributions from processing speed, item identification, and short-term memory. *Journal of Experimental Child Psychology*, 65, 1–24.
- Butterfield, E. C., & Ferretti, R. P. (1987). Toward a theoretical integration of cognitive hypotheses about intellectual differences among children. In J. G. Borkowski & J. D. Day (Eds.), *Cognition in special children: Comparative approaches to retardation, learning disabilities, and giftedness* (pp. 195–233). Norwood, NJ: Ablex.
- Carpenter, T. P., & Moser, J. M. (1984). The acquisition of addition and subtraction concepts in grades one through three. *Journal of Research in Mathematics Education*, 15, 179–203.
- Case, L. P., Harris, K. R., & Graham, S. (1992). Improving the mathematics problem-solving skills for students with learning disabilities. *Journal of Learning Disabilities*, 30, 130–141.
- Case, R. (1985). *Intellectual development: Birth to adulthood*. San Diego, CA: Academic Press.
- Cirino, P. T., Ewing-Cobbs, L., Barnes, M., Fuchs, L. S., & Fletcher, J. M. (2007). Cognitive arithmetic differences in learning disability groups and the role of behavioral inattention. *Learning Disabilities Research and Practice*, 22, 25–35.
- CTB/McGraw-Hill. (1997). *TerraNova technical manual*. Monterey, CA: Author.
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19, 450–466.
- Fletcher, J. M., Shaywitz, S. E., Shankweiler, D. P., Katz, L., Liberman, I. Y., Stuebing, K. K., et al. (1994). Cognitive profiles of reading disability: Comparisons of discrepancy and low achievement definitions. *Journal of Educational Psychology*, 85, 1–18.
- Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology*, 97, 493–513.
- Fuchs, L. S., & Fuchs, D. (2002). Mathematical problem-solving profiles of students with mathematics disabilities with and without comorbid reading disabilities. *Journal of Learning Disabilities*, 35, 563–573.



- Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., et al. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology*, 98, 29–43.
- Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1990). *Test of Computational Fluency*. (Available from L. S. Fuchs, 328 Peabody Vanderbilt University, Nashville, TN 37203)
- Fuchs, L. S., Hamlett, C. L., & Powell, S. R. (2003). *Grade 3 Math Battery*. (Available from L. S. Fuchs, 328 Peabody, Vanderbilt University, Nashville, TN 37203)
- Geary, D. C. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin*, 114, 345–362.
- Geary, D. C., Brown, S. C., & Samaranayake, V. A. (1991). Cognitive addition: A short longitudinal study of strategy choice and speed-of-processing differences in normal and mathematically disabled children. *Developmental Psychology*, 27, 787–797.
- Geary, D. C., Widaman, K. F., Little, T. D., & Cormier, P. (1987). Cognitive addition: Comparison of learning disabled and academically normal elementary school children. *Cognitive Development*, 2, 249–269.
- Hanich, L. B., Jordan, N. C., Kaplan, D., & Dick, J. (2001). Performance across different areas of mathematical cognition in children with learning disabilities. *Journal of Educational Psychology*, 93, 615–626.
- Harris, R. J. (1975). *A primer of multivariate statistics*. San Diego, CA: Academic Press.
- Hecht, S. A., Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2001). The relations between phonological processing abilities and emerging individual differences in mathematical computational skills: A longitudinal study from second to fifth grades. *Journal of Experimental Child Psychology*, 79, 192–227.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Hitch, G. J., & McAuley, E. (1991). Working memory in children with specific arithmetical learning disabilities. *British Journal of Psychology*, 82, 375–386.
- Huberty, C. J. (1975). Discriminant analysis. *Review of Educational Research*, 45, 543–598.
- Jordan, N. C., & Hanich, L. B. (2000). Mathematical thinking in second-grade children with different forms of LD. *Journal of Learning Disabilities*, 33, 567–578.
- Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). Arithmetic fact mastery in young children: A longitudinal investigation. *Journal of Experimental Child Psychology*, 85, 103–119.
- Jordan, N. C., Levine, S. C., & Huttenlocher, J. (1995). Calculation abilities in young children with different patterns of cognitive functioning. *Journal of Learning Disabilities*, 28, 53–64.
- Kintsch, W., & Greeno, J. G. (1985). Understanding and solving word arithmetic problems. *Psychological Review*, 92, 109–129.
- Landerl, K., Began, A., & Butterworth, B. (2004). Developmental dyscalculia and basic numerical capacities: A study of 8–9-year-old students. *Cognition*, 93, 99–125.
- LeBlanc, M. D., & Weber-Russell, S. (1996). Text integration and mathematics connections: A computer model of arithmetic work problem-solving. *Cognitive Science*, 20, 357–407.
- Lemaire, P., & Siegler, R. S. (1995). Four aspects of strategic change: Contributions to children's learning of multiplication. *Journal of Experimental Psychology: General*, 124, 83–97.
- Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenger, O. (2006). *SAS system for mixed models* (2nd ed.). Cary, NC: SAS Institute.
- Luria, A. R. (1980). *Higher cortical functions in man* (2nd ed.). New York: Basic Books.
- Mayer, R. E. (1992). *Thinking, problem solving, cognition* (2nd ed.). New York: Freeman.
- Newcomer, P. L., & Hammill, D. D. (1988). *Test of Language Development* (Rev. ed.). Austin, TX: PRO-ED.
- Passolunghi, M. C., Cornoldi, C., & De Liberto, S. (1999). Working memory and intrusions of irrelevant information in a group of specific poor problem solvers. *Memory and Cognition*, 27, 779–790.
- Passolunghi, M. C., & Siegel, L. S. (2001). Short-term memory, working memory, and inhibitory control in children with specific arithmetic learning disabilities. *Journal of Experimental Child Psychology*, 80, 44–57.
- Passolunghi, M. C., & Siegel, L. S. (2004). Working memory and access to numerical information in children with disability in mathematics. *Journal of Experimental Child Psychology*, 88, 348–367.
- Perfetti, C. A. (1992). The representation problem in reading acquisition. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 145–174). Hillsdale, NJ: Erlbaum.
- Pickering, S., & Gathercole, S. (2001). *Working Memory Test Battery for Children*. London: Psychological Corporation.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.
- Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, 5, 199–213.
- Riley, M. S., & Greeno, J. G. (1988). Developmental analysis of understanding language about quantities and of solving problems. *Cognition and Instruction*, 5, 49–101.
- Riley, M. S., Greeno, J. G., & Heller, J. I. (1983). Development of children's problem-solving ability in arithmetic. In *The development of mathematical thinking* (pp. 153–196). San Diego, CA: Academic Press.
- Rourke, B. P., & Finlayson, M. A. J. (1978). Neuropsychological significance of variations in patterns of academic skills: Verbal and visual-spatial abilities. *Journal of Abnormal Child Psychology*, 6, 121–133.
- Russell, R. L., & Ginsburg, H. P. (1984). Cognitive analysis of children's mathematical difficulties. *Cognition and Instruction*, 1, 217–244.
- Seaman, M. A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practical problems. *Psychological Bulletin*, 110, 577–586.
- Siegel, L. S., & Linder, B. (1984). Short-term memory process in children with reading and arithmetic disabilities. *Developmental Psychology*, 20, 200–207.
- Siegel, L. S., & Ryan, E. B. (1989). The development of working memory in normally achieving and subtypes of learning disabled children. *Child Development*, 60, 973–980.
- Swanson, H. L. (1993). An information processing analysis of learning disabled children's problem solving. *American Educational Research Journal*, 30, 861–893.
- Swanson, H. L. (2006). Cross-sectional and incremental changes in working memory and mathematical problem solving. *Journal of Educational Psychology*, 98, 265–281.
- Swanson, H. L., Ashbaker, M., & Sachse-Lee, C. (1996). Learning disabled readers' working memory as a function of processing demands. *Journal of Experimental Psychology*, 61, 242–275.
- Swanson, H. L., & Beebe-Frankenberger, M. (2004). The relationship between working memory and mathematical problem-solving in children at risk and not at risk for serious math difficulties. *Journal of Educational Psychology*, 96, 471–491.
- Swanson, H. L., Cooney, J. B., & Brock, S. (1993). The influence of working memory and classification ability on children's word problem solution. *Journal of Experimental Child Psychology*, 55, 374–395.
- Swanson, H. L., & Sachse-Lee, C. (2001). Mathematical problem-solving and working memory in children with learning disabilities: Both executive and phonological processes are important. *Journal of Experimental Child Psychology*, 79, 294–321.
- Webster, R. E. (1979). Visual and aural short-term memory capacity

deficits in mathematics disabled students. *Journal of Educational Research*, 72, 272-283.

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence*. San Antonio, TX: Psychological Corporation.

Wilson, K. M., & Swanson, H. L. (2001). Are mathematics disabilities due to a domain-general or a domain-specific working memory deficit? *Journal of Learning Disabilities*, 34, 237-248.

Woodcock, R. W. (1997). *Woodcock Diagnostic Reading Battery*. Itasca, IL: Riverside.

Woodcock, R. W. (1998). *Woodcock Reading Mastery Tests—Revised*. Circle Pines, MN: American Guidance Service.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-Educational Battery—Revised*. Allen, TX: DLM Teaching Resources.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement*. Itasca, IL: Riverside.

Zhu, J. (1999). *WASI manual*. San Antonio, TX: Psychological Corporation.

Received December 5, 2006

Revision received August 11, 2007

Accepted August 13, 2007 ■



## AMERICAN PSYCHOLOGICAL ASSOCIATION SUBSCRIPTION CLAIMS INFORMATION

Today's Date: \_\_\_\_\_

We provide this form to assist members, institutions, and nonmember individuals with any subscription problems. With the appropriate information we can begin a resolution. If you use the services of an agent, please do **NOT** duplicate claims through them and directly to us. **PLEASE PRINT CLEARLY AND IN INK IF POSSIBLE.**

PRINT FULL NAME OR KEY NAME OF INSTITUTION \_\_\_\_\_

MEMBER OR CUSTOMER NUMBER (MAY BE FOUND ON ANY PAST ISSUE LABEL) \_\_\_\_\_

ADDRESS \_\_\_\_\_

DATE YOUR ORDER WAS MAILED (OR PHONED) \_\_\_\_\_

CITY \_\_\_\_\_ STATE/COUNTRY \_\_\_\_\_ ZIP \_\_\_\_\_

PREPAID \_\_\_\_\_ CHECK \_\_\_\_\_ CHARGE  
CHECK/CARD CLEARED DATE: \_\_\_\_\_

YOUR NAME AND PHONE NUMBER \_\_\_\_\_

(If possible, send a copy, front and back, of your cancelled check to help us in our research of your claim.)

ISSUES: \_\_\_\_\_ MISSING \_\_\_\_\_ DAMAGED

TITLE \_\_\_\_\_

VOLUME OR YEAR \_\_\_\_\_

NUMBER OR MONTH \_\_\_\_\_

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

\_\_\_\_\_  
\_\_\_\_\_  
\_\_\_\_\_

*Thank you. Once a claim is received and resolved, delivery of replacement issues routinely takes 4-6 weeks.*

(TO BE FILLED OUT BY APA STAFF)

DATE RECEIVED: \_\_\_\_\_  
ACTION TAKEN: \_\_\_\_\_  
STAFF NAME: \_\_\_\_\_

DATE OF ACTION: \_\_\_\_\_  
INV. NO. & DATE: \_\_\_\_\_  
LABEL NO. & DATE: \_\_\_\_\_

Send this form to APA Subscription Claims, 750 First Street, NE, Washington, DC 20002-4242

**PLEASE DO NOT REMOVE. A PHOTOCOPY MAY BE USED.**



# Mathematical Competencies in Children With Different Types of Learning Difficulties

Ulf Andersson  
Linköping University

The mathematical performance of 182 third and fourth graders in 8 different areas of mathematics was examined. The children belonged to 4 achievement groups: children with mathematic difficulties (MD only), children with both mathematic and reading difficulties (MD–RD), children with reading difficulties (RD only), and normally achieving children (control group). Both MD groups performed worse than the normally achieving children in all but 1 area, place value knowledge. The MD-only and the MD–RD children performed equally in all areas of mathematics. The RD-only group performed at the same level as the control group on all areas of mathematics. The study provides further evidence that fact retrieval deficits are a cardinal characteristic of children with MD. The MD children's substantial difficulties with mathematic word problem solving can be attributed to several processes involved in problem solving. Besides poor skills in multidigit calculation, arithmetic fact retrieval, and poor understanding of calculation principles, children with MD might have deficits related to specific problem-solving processes such as establishing a problem representation and developing a solution plan.

**Keywords:** mathematical difficulties, reading difficulties, calculation, mathematic problem solving, telling time

During the past 10 to 15 years, an increasing number of studies have been carried out on children with difficulties in mathematics (Geary & Brown, 1991; Geary, Brown, & Samaranayake, 1991; Gonzalez & Espinel, 2002; Russell & Ginsburg, 1984; Swanson & Beebe-Frankenberger, 2004). However, most of these studies have focused on basic mathematical skills (i.e., arithmetic fact retrieval, calculation skill, and simple arithmetic word problem solving; Geary, Hamson, & Hoard, 2000; Geary, Hoard, & Hamson, 1999; Ostad, 1997, 1998), whereas considerably less research has focused on other domains of mathematics such as approximate arithmetic, comprehension of calculation principles, and mathematical word problem solving (Fuchs & Fuchs, 2002; Hanich, Jordan, Kaplan, & Dick, 2001; Russell & Ginsburg, 1984).

The present study was an attempt to extend existing research by examining multiple areas of mathematics in groups of children with different types of learning difficulties. Up to now, only four studies have examined multiple areas of mathematics in groups of children with learning difficulties (Hanich et al., 2001; Jordan & Hanich, 2000; Jordan, Hanich, & Kaplan, 2003; Russell & Ginsburg, 1984). This type of research is necessary if one is to be able to provide a more comprehensive picture of mathematical difficulties (MD) in children, for instance by identifying core deficits. It is quite possible that children with MD might display normal skill levels in some areas of mathematics, whereas their skill levels

in other areas are severely impaired (cf. Dowker, 2005; Ginsburg, 1997). By examining multiple areas of mathematics, it should also be possible to study whether difficulties in one area are due to problems in another area; for example, problems with multidigit calculation might be attributed to deficits in arithmetic fact retrieval.

## Different Domains of Mathematics

The general theoretical assumption of the present study was that mathematical competence consists of multiple abilities that are taught and learned in a hierarchical manner (Aunola, Leskinen, Lerkkanen, & Nurmi, 2004; Dowker, 2005; Geary, 1994). That is, basic skills such as more/less comparisons of small quantities and counting are prerequisites for solving basic arithmetic tasks (e.g.,  $3 + 4 = 7$ ), first by means of counting procedures and later by direct retrieval of arithmetic facts from long-term memory (Baroody & Wilkins, 1999; Geary et al., 2000; Mazzocco & Thompson, 2005). More complex mathematic skills such as multidigit calculation and word problem solving are in turn facilitated by mastery of basic arithmetic operation; arithmetic facts retrieval; and conceptual understanding of place value, the base-10 number system, and calculation principles (Geary, 2004; Jordan, Hanich, & Uberti, 2003).

The present study examined eight areas of mathematics: arithmetic fact retrieval, written arithmetic calculation, approximate arithmetic, place value, calculation principles, one-step and multistep mathematic word problems, and telling time. These areas were selected because they have either direct or indirect relevance to the teaching of mathematics in the present age group of children (i.e., Grades 3 and 4). That is, all areas are emphasized in elementary school, with a particular focus on basic arithmetic (including arithmetic fact retrieval), place value, and written calculation in

---

Ulf Andersson, Department of Behavioural Sciences and Learning, Linköping University, Linköping, Sweden.

This research was supported by Grant Dnr 2003-0158 from the Swedish Council for Working Life and Social Research awarded to Ulf Andersson.

Correspondence concerning this article should be addressed to Ulf Andersson, Department of Behavioural Sciences and Learning, Linköping University, SE-581 83 Linköping, Sweden. E-mail: ulfan@ibv.liu.se

Grades 1 and 2, and more focus on mathematic word problem solving in Grades 3 and 4. Thus, skills in areas such as fact retrieval and place value should be rather well developed in most children, whereas the ability to solve word problems should still be developing. Another reason for examining the present areas was that previous studies have shown that children in Grades 2 to 4 with MD display problems with almost all of these areas (Hanich et al., 2001; Jordan & Hanich, 2000; Jordan, Hanich, & Kaplan, 2003; Jordan & Montani, 1997).

### Children With Specific Difficulties in Mathematics and Comorbid Reading Difficulties

When performing research on children with MD it is important to make a distinction between children with specific MD (MD only) and children with difficulties in both mathematics and reading, so-called comorbid mathematic and reading difficulties (MD–RD). This distinction is important because the functional and general cognitive characteristics of children with MD only and comorbid MD–RD differ (Andersson & Lyxell, 2007; Geary et al., 1991, 1999; Siegel & Ryan, 1989). Overall, children with comorbid MD–RD display more severe and global functional difficulties than children with MD only (Andersson & Lyxell, 2007; Fuchs & Fuchs, 2002; Jordan & Hanich, 2000; Jordan & Montani, 1997). Research has also indicated that children with MD–RD have a general working-memory deficit, whereas children with MD only have a specific working-memory deficit (Geary et al., 1991, 1999; Siegel & Ryan, 1989).

### Basic Arithmetic

Previous research focusing on basic arithmetic skills has shown that children with MD only and with comorbid MD–RD are slower, commit more errors, and employ developmentally immature counting procedures (e.g., counting all instead of counting on) when solving simple arithmetic problems ( $3 + 4$ ) by means of different counting strategies such as finger counting, verbal counting, or silent counting (Geary et al., 1999, 2000; Hanich et al., 2001; Russell & Ginsburg, 1984; see also Geary, 1993, 2004, for a review). Moreover, they do not follow the normal development of shifting from procedural problem-solving strategies (e.g., silent counting) to a memory-based problem-solving strategy. That is, they solve simple arithmetic problems ( $3 + 4$ ) by means of direct and automatic retrieval of arithmetic facts from long-term memory less frequently compared to children without MD. They often continue to employ procedural problem-solving strategies as their main solution strategies, and when they apply direct retrieval they are slow and commit many errors (Bull & Johnston, 1997; Geary et al., 1991; Jordan & Montani, 1997; Ostad, 1997, 1998; Siegler, 1988). These findings indicate that children with MD and MD–RD have procedural problems and problems with establishing memory representations regarding the associations between problems and answers to simple arithmetic problems in semantic long-term memory (Ashcraft, 1992; Geary, 1993). Lack of these representations, or difficulties in retrieving them, could prevent the child from developing the automation that is necessary in order to effectively solve complex (multidigit) arithmetic tasks (Ashcraft, 1992). This is because lack of automation will probably require more of the child's limited working-memory capacity, leaving less

capacity available for performing other aspects of arithmetic calculation, such as calculating and storing interim results and performing carrying or borrowing operations (Ashcraft, 1992, 1995; Geary, 2004).

Fact retrieval problems have been demonstrated from Grade 1 up to Grade 7 in children with MD, thus it seems that fact retrieval deficits are persistent and a cardinal characteristic of MD in children (Geary et al., 1999; Jordan, Hanich, & Kaplan, 2003; Ostad, 1997, 1998; Russell & Ginsburg, 1984). Although children with both MD only and MD–RD have fact retrieval problems and rely on backup strategies to a greater extent than normally achieving children, the MD-only group uses counting procedures more skillfully than the MD–RD children (Andersson & Lyxell, 2007; Geary et al., 2000; Hanich et al., 2001; Jordan & Montani, 1997). This indicates that the MD–RD group has more severe procedural problems preventing them from obtaining correct answers to simple arithmetic combinations by executing counting strategies (e.g., verbal counting). The more severe procedural problems in children with MD–RD might be related to a less developed conceptual understanding of counting principles (Geary, Bow-Thomas, & Yao, 1992; Geary et al., 1999). As working-memory resources support counting processes, it is also possible that the more global working-memory problems of the MD–RD children account for their problem with solving arithmetic combinations by means of verbal counting strategies (Geary, 1993, 2004; Geary et al., 1999; Logie & Baddeley, 1987; Passolunghi & Siegel, 2001).

In sum, a limitation of the studies described previously is that they almost exclusively have employed addition problems and to some extent subtraction problems, but not multiplication problems. In order to expand researchers' knowledge concerning fact retrieval problems in children with MD, the ability to retrieve multiplication facts should be examined, as children's problems might vary among operations.

### Mathematic Story Problem Solving

Most studies examining word problem solving in children with MD have used simple one-step arithmetic problems such as "Pete has 7 marbles. Sam gives him 5 more marbles. How many marbles does Pete have now?" or "Kate and Mark are playing marbles. Kate has 2 marbles. Mark gives her some more marbles. Now Kate has 6 marbles. How many marbles did Mark give her?" (e.g., Hanich et al., 2001; Jordan & Hanich, 2000; Jordan, Hanich, & Kaplan, 2003; Jordan & Montani, 1997). The empirical picture of these studies demonstrates that children with MD or MD–RD in Grades 2, 3, and 4 have major difficulties with this type of word problem-solving task. However, children with MD–RD appear to have more severe difficulties in solving simple one-step arithmetic problems than children with MD only (e.g., Hanich et al., 2001; Jordan & Hanich, 2000; Jordan, Hanich, & Kaplan, 2003; Jordan & Montani, 1997). For example, Jordan and Montani found that children with MD only obtained lower scores than normally achieving children in timed conditions but not in untimed conditions, whereas children with MD–RD obtained lower scores regardless of whether the task was timed or untimed compared to normally achieving children. Furthermore, in the untimed condition, MD–RD children were outperformed by MD-only children. According to Jordan and colleagues, these findings indicate that children with MD only have problems with problem-solving



speed, whereas MD–RD children seem to have difficulties with problem comprehension (Hanich et al., 2001; Jordan & Hanich, 2000; Jordan & Montani, 1997). This conclusion is supported by findings reported by Fuchs and Fuchs (2002).

A few researchers have also examined problem solving beyond simple arithmetic problems (Fuchs & Fuchs, 2002; Parmar, Cawley, & Frazita, 1996; Russell & Ginsburg, 1984). Russell and Ginsburg found that children with MD–RD performed on a par with controls on simple one-step problems but displayed considerable difficulties with multistep word problems and problems that involved irrelevant information. Parmar et al. reported similar findings from a study on children with mild forms of MD–RD across Grades 3 through 8. Fuchs and Fuchs compared the performance of fourth graders with MD only or MD–RD to normative data on three types of problem-solving tasks: one-step arithmetic problems, multistep complex story problems that included irrelevant information, and real-world story problems that included irrelevant information. They also separated operational ability and problem-solving ability by scoring on each dimension. Both groups with MD displayed large deficits on all three tasks, but the MD-only children performed overall better than the MD–RD children. However, on the arithmetic story problems the MD-only group outperformed the MD–RD group on both dimensions (operations and problem solving), whereas the two groups performed equally on the operation dimension on the complex and real-world problem task. In contrast, the MD-only group outperformed the MD–RD group on the problem-solving dimension on the two more difficult problem-solving tasks. Fuchs and Fuchs posited that these findings were due to the fact that the complex story problem and real-world problem tasks involved more difficult operational demands than the arithmetic story problem task. The latter task required only number fact skills, whereas the two other tasks required algorithm and application skills as well as number fact skills. Thus, the advantage of the MD-only children over the MD–RD children concerning operational ability was restricted to simple number fact operation but did not involve more complex mathematical operations. The MD-only children, in contrast, appeared to have overall better developed problem comprehension abilities (establishing a problem representation and a solution plan; Kintsch & Greeno, 1985; Mayer & Hegarty, 1996) than children with MD–RD (Fuchs & Fuchs, 2002; see also Hanich et al., 2001; Jordan & Hanich, 2000; Jordan & Montani, 1997).

In sum, previous research has provided consistent evidence that arithmetic and mathematic problem solving is a mathematic domain that imposes severe difficulties for children with MD, and especially for children with comorbid RD. However, more research is required that focuses on complex mathematic word problem-solving ability in children with MD and MD–RD. The more severe comprehension and problem-solving difficulties in children with MD–RD are most likely, to some extent, a result of their reading deficit, which might prevent the children from having full access to the problem text. It is also possible that the reading deficit in children with MD–RD reflects a general weakness in language comprehension. As such, it would constitute a major obstacle when the child must comprehend word problems even if the problem is orally presented to the child.

## Place Value and Written Multidigit Calculation

Only four studies have included measures tapping children's understanding of place value and skill in written multidigit calculation when studying children with MD. This is surprising considering that place value is a basic mathematic concept, and written multidigit calculation can be considered a natural second step (after basic number fact skills) in children's learning of basic mathematic skills. In all four available studies children with MD–RD obtained significantly lower scores on multidigit calculation tasks than normally achieving children (Hanich et al., 2001; Jordan & Hanich, 2000; Jordan, Hanich, & Kaplan, 2003; Russell & Ginsburg, 1984). Furthermore, Jordan and Hanich and Hanich et al. found that children with RD only performed better than MD–RD children. Jordan and colleagues also found that second and third graders with MD–RD had a poorer understanding of place value compared to normally achieving children (Hanich et al., 2001; Jordan & Hanich, 2000; Jordan, Hanich, & Kaplan, 2003). In Russell and Ginsburg's study of fourth graders, children with MD–RD displayed as good an understanding of place value as the age-matched controls. In relation to children with MD only, two studies have demonstrated that they have a weakness in written multidigit calculation and poorer understanding of place value in comparison to age-matched controls (Hanich et al., 2001; Jordan, Hanich, & Kaplan, 2003). Jordan and Hanich, in contrast, found that children with MD only performed on a par with normally achieving children but better than MD–RD children on written multidigit calculation. An error analysis indicated that the performance pattern of MD–RD children was qualitatively different from that of children with MD only. The MD–RD children's errors indicated that they have poor reasoning or judgment skills, whereas the MD-only children's errors consisted of calculation bugs (e.g., minor miscalculations, wrong operations, etc.) similar to those made by normally achieving children.

In sum, due to the limited research no firm conclusions can be drawn with regard to these aspects of mathematic cognition and MD. The only conclusion that can be drawn with some confidence is that children with MD have problems with multidigit calculation. The concept of place value might be a problem for second and third graders with MD–RD but not fourth graders. Previous studies have also given some indication that poor understanding of place value might have a negative effect on the child's ability to perform written multidigit calculations (Jordan, Hanich, & Kaplan, 2003).

## Calculation Principles

Developing an understanding of the relationships within and between arithmetic operations (i.e., calculation principles) is a prerequisite for accurate and efficient multidigit calculation and mathematic word problem solving (Geary, 1994; Jordan, Hanich, & Uberty, 2003). Studies performed by Jordan and colleagues (Hanich et al., 2001; Jordan, Hanich, & Kaplan, 2003) have indicated that both groups of children with MD have a weakness regarding this aspect of mathematics in Grades 2 and 3. Jordan, Hanich, and Kaplan (2003) also found that third graders with MD–RD have a poorer understanding of calculation principles compared to children with MD only. Russell and Ginsburg (1984), however, did not find any evidence that fourth graders with MD–RD have a vague understanding of calculation principles. So,



the empirical picture concerning the understanding of calculation principles in children with MD is unclear due to there being few available studies. The present study, therefore, sought to examine this area of mathematics in children with MD.

### Approximate Arithmetic

Findings reported by Hanich et al. (2001) and Jordan, Hanich, and Kaplan (2003) have suggested that approximate arithmetic (i.e., the ability to quickly provide an estimated result of an arithmetic problem) is a core deficit for children with MD only and MD-RD. Brain imaging data suggest that when performing this type of task, the individual employs a mental spatial number line that draws upon visuospatial circuits of the dorsal parietal pathway of the brain (Dehaene, Spelke, Pinel, Stanescu, & Tsivkin, 1999; see also Dehaene & Cohen, 1991). Thus, approximate arithmetic seems to rely on visuospatial abilities and seems to be independent of language (Dehaene et al., 1999; Hanich et al., 2001; Jordan, Hanich, & Kaplan, 2003). The empirical picture regarding approximate arithmetic is, however, far from clear as only two studies have provided evidence that children with MD have a weakness within this domain. To further complicate the picture, Russell and Ginsburg (1984) found that children in Grade 4 with MD-RD did not perform worse than normally achieving children. Hence, more research is required to obtain a clearer view of the relationship between MD and approximate arithmetic.

### Time Telling

Time telling is a complex and important cognitive skill that is required on a daily basis (Bock, Irwin, Davidson, & Levelt, 2003; Friedman & Laycock, 1989). Available studies have shown that the ability to tell time from analog clocks develops in a particular order. Most children are able to tell whole-hour times by age 6, half-hour times by age 7, quarter-hour times by age 9, and 5- and 1-min times by ages 8 to 10 (Case, Sandieson, & Dennis, 1986; Friedman & Laycock, 1989; Siegler & McGilly, 1989; Vakali, 1991). According to Friedman and Laycock's study, the ability to tell the time of digital clocks develops much earlier than that of analog clocks. The researchers found that all second graders (mean age = 7.70 years) were able to correctly identify whole-hour times, half-hour times, and 1-min times for digital clocks. The similar performance level for analog clocks was not reached until age 10.

Previous research has also demonstrated that children employ a mixture of retrieval and procedural strategies when they tell time, and that older children rely more on retrieval strategies than younger children (Case et al., 1986; Friedman & Laycock, 1989; Siegler & McGilly, 1989; Vakali, 1991). According to Friedman and Laycock (1989), time telling

includes attending to the numerals to which the hour hand points or, if necessary, has passed most recently to get the hour value, determining the minute value by counting clockwise by 5s from the top of the clock face until the 5-min mark pointed to or immediately preceding the minute hand pointing is reached, and, if necessary, counting the remaining hash marks by 1s. (p. 357)

Furthermore, time-telling problems such as whole-hour analog times and half-hour analog times seem to be solved by retrieving

a preexisting answer from long-term memory. Thus, it is assumed that the child develops memory representations regarding the associations between particular configurations and the time names (Friedman & Laycock, 1989; Siegler & McGilly, 1989). Results reported by Friedman and Laycock have indicated that telling the time of digital clocks seems to involve quite different processes compared to that of analog clocks. Telling the time of digital clocks (e.g., 13:25) is quite similar to decoding regular one- and two-digit numbers (9 or 22), as the hour and minute values are presented as numbers. Thus, time telling of digital clocks is basically a question of retrieving number names. However, the child needs to know that the colon is used to separate the hour value and the minute value, that X:00 is read as "o'clock," and that a zero in the tens place to the right of the colon is read as "oh." Because telling time is a skill that in many respects resembles basic arithmetic (counting procedures and fact retrieval), it would be interesting to examine the time-telling skills in children with MD only and MD-RD, something that was not done in previous studies.

### Mathematic Skill in Children With RD Only

MD in children often coexist with RD (Ackerman & Dykman, 1995; Gathercole, Alloway, Willis, & Adams, 2006; Geary, 1993; Lewis, Hitch, & Walker, 1994). Reading ability also seems to have an impact on children's mathematical ability, as children with MD-RD perform worse than children with MD only within specific areas of mathematics (e.g., story problems and exact calculation of arithmetic combinations), whereas children with MD-RD display similar levels of reading ability as children with RD only. As a matter of fact, Jordan, Kaplan, and Hanich (2002) presented evidence that suggests that children identified with RD only in early grades (second grade) are at risk for developing MD in higher grades (third grade). Consistent with the assumption that the RD in children with RD only may have a negative impact on these children's mathematic skills, Geary et al. (2000) and Hanich et al. (2001) observed that children with RD only achieved significantly worse results than age-matched controls on a forced addition fact retrieval task. Jordan and colleagues (Jordan & Hanich, 2000; Jordan, Hanich, & Kaplan, 2003) demonstrated that second and third graders with RD only may have a vague understanding of the concept of place value. Furthermore, Jordan, Hanich, and Kaplan (2003) found that third graders with RD only obtained lower scores on written multidigit calculation and a simple story problem task. Thus, in line with the findings of Jordan et al. (2002), it is possible that children with RD only may encounter difficulties in later elementary school within certain areas of mathematics (e.g., complex word problem solving) that are mediated via language (i.e., reading). It should, however, be pointed out that previous studies on children with RD only have shown that these children perform on a par with normally achieving children on most aspects of mathematic cognition (Geary et al., 1999; Jordan & Hanich, 2000). Thus, it remains to be seen whether children with RD only have weaknesses in specific domains of mathematics.

To summarize, research has shown that children with MD or MD-RD have weaknesses in many areas of mathematics, but it is important to regard children with MD only and children with MD-RD as separate subgroups of children with MD. Overall, children with MD-RD display more severe difficulties than children with MD only. More specifically, children with MD-RD have



greater deficiencies with arithmetic problem solving and exact calculation of arithmetic combinations, whereas their problems in other areas of mathematic cognition (such as forced retrieval of arithmetic facts, approximate arithmetic, place value, calculation principles, and written calculation) are equal to those of children with MD only. So, children with MD only show an advantage over children with MD–RD in areas that can be mediated by language (e.g., story problems and exact calculation of arithmetic combinations) but not in areas that appear to depend on numerical magnitude and automaticity (e.g., approximate arithmetic, arithmetic fact retrieval). The negative influence RD have on the mathematical ability of children with MD–RD might also be present for children with RD only; that is, these children may encounter difficulties in later elementary school with areas of mathematics that are mediated via language (i.e., reading).

### The Present Study

The present study was an attempt to extend previous findings regarding mathematic competencies of children with MD only, MD–RD, and RD only. To reach this goal, relatively large samples of children with MD only, MD–RD, and also RD only were assessed on multiple areas of mathematic cognition and compared to children with normal achievement. Similar to studies by Jordan and colleagues (Hanich et al., 2001; Jordan, Hanich, & Kaplan, 2003; see also Russell & Ginsburg, 1984), measures of addition and subtraction fact retrieval, written multidigit calculation, calculation principles, place value, and approximate arithmetic were employed. In addition to these measures, a multiplication fact retrieval task was included in the study. In contrast to the majority of previous studies, the present study included one-step mathematic word problems that required multidigit calculation as well as number fact skills and complex mathematic word problems that required multistep solutions and multidigit calculation (cf. Fuchs & Fuchs, 2002). As time telling is an important skill that is required on a daily basis, the present study assessed this aspect of mathematics in the four subgroups of children (including normally achieving children). Screening tests of mathematics and reading were included to establish that the groups of children classified as having MD and/or RD indeed performed significantly poorer than the group of normally achieving children.

### Supporting Cognitive Functions and Measures of IQ

A large body of research has demonstrated that many areas of mathematic cognition in children are supported by basic cognitive functions such as working-memory capacity, general processing speed, and executive functions (Andersson, in press; Bull & Scerif, 2001; Geary, 2004; Passolunghi & Pazzaglia, 2004; Swanson, 1994, 2006; Swanson & Beebe-Frankenberger, 2004). In addition, these basic cognitive functions have been found to be poorer in children with MD and MD–RD compared to age-matched normally achieving children (Andersson & Lyxell, 2007; Bull & Johnston, 1997; Hitch & McAuley, 1991; McLean & Hitch, 1999; Passolunghi & Siegel, 2004; Siegel & Ryan, 1989; Swanson, 1994; Swanson & Beebe-Frankenberger, 2004). Hence, in order to examine and, if necessary, control for possible group differences in basic cognitive functioning, three tasks tapping these functions were included in the study (visual-matrix span, number matching,

trail making; Andersson & Lyxell, 2007; Bull & Johnston, 1997). Raven's progressive matrices (Raven, 1976) and a verbal ability task (Järpsten & Taube, 1997) were also employed in order to examine group differences in mathematics beyond the influence of verbal and nonverbal IQ.

In sum, by using larger samples and more extensive measurement than previous research, the present study has the potential to contribute to the research literature on MD in children. The present article examined data from the first-year assessment of a 3-year longitudinal study on (a) the development of mathematic skills in children with MD only, RD only, and MD–RD; and (b) how development in mathematics is connected to reading development and vice versa. However, the reading-related tasks included in this project are not presented, because they do not provide any additional information to the present article that focuses on different areas of mathematic cognition. The reading tasks will hopefully provide important information when data have been collected for all 3 years.

### Predictions and Specific Research Questions

The following predictions and specific research questions were examined in relation to the different areas of mathematic cognition.

1. It was predicted that children with MD only and MD–RD would show weaknesses in all of the examined areas of mathematic cognition. That is, they would perform significantly poorer than the normally achieving children on all mathematic tasks. It was, however, expected that both groups of children with MD would display more severe problems with arithmetic facts retrieval, mathematic word problem solving, and written multidigit arithmetic calculation compared to areas such as place value and calculation principles (cf. Hanich et al., 2001; Jordan & Hanich, 2000; Jordan, Hanich, & Kaplan, 2003; Russell & Ginsburg, 1984).

2. The two groups with MD were expected to perform equally on forced arithmetic fact retrieval, approximate arithmetic, place value, calculation principles, and written calculation. However, as mathematic word problem solving is mediated by language, the MD-only children were expected to have an advantage over the MD–RD children in word problem solving due to their better developed reading (language) skills and also due to their better problem comprehension abilities (cf. Fuchs & Fuchs, 2002; Hanich et al., 2001; Jordan & Hanich, 2000; Jordan & Montani, 1997).

3. A specific interest of the present study was to examine whether weaknesses in arithmetic fact retrieval can account for the expected low performance of the two MD groups on written calculation. That is, will children with MD only and MD–RD still perform lower than the controls on written multidigit calculation when expected group differences in arithmetic fact retrieval are controlled for?

4. As the understanding of calculation principles and skills in written calculation is important for accurate and efficient problem solving, the present study sought to examine whether the poor performance of the MD-only group and the MD–RD group on mathematic word problem solving is due to the previously mentioned mathematic skill. That is, will children with MD only and MD–RD still perform lower than the controls on mathematic word

problem solving when expected group differences in written multidigit calculation and understanding of calculation principles are controlled for?

5. A specific interest of the present study was to examine whether children with RD only have weaknesses in specific domains of mathematics that are mediated by language (i.e., word problem solving, arithmetic facts, written multidigit calculation; Dehaene, 1992; Geary, 1993, 2004; Kulak, 1993).

6. As telling time is a skill that in many respects resembles basic arithmetic (counting procedures and fact retrieval), the present study also sought to address the question whether children with MD only or MD–RD have difficulties with this aspect of mathematics.

## Method

### Participants

A total of 182 third ( $n = 88$ ) and fourth graders ( $n = 94$ ) attending 28 schools in the southern parts of Sweden participated in this study. The children were recruited by means of a letter of consent that the children took home to the parents from school. Fortunately, very few of the children with MD only, RD only, and MD–RD who were asked to participate in the study declined to participate, and because the children were drawn from many schools (i.e., 28) one might assume that the present children provide representative samples of the populations of children with MD only, RD only, and MD–RD. The total sample had a mean age of 125 months ( $SD = 7.00$ ). All children were fluent speakers of Swedish, had normal or corrected-to-normal visual acuity, and had no hearing loss. Children diagnosed with neurological disturbances (e.g., attention-deficit/hyperactivity disorder) were excluded from the study.

The criterion for being classified into the three disability groups was that the child received special instruction either in mathematics (MD only), reading (RD only), or both mathematics and reading (MD–RD) at the time of the study. An additional selection criterion was that the child's score on Raven's Standard Progressive Matrices test and the verbal ability test was at most 1.50  $SD$  below the group mean of the normally achieving children of the same age. The decision to provide special instruction to the children with MD only, RD only, or MD–RD in the present study was based on several sources of information. Potential learning problems in relation to mathematics and reading were first observed by the classroom teachers. The teachers' observation of the children's poor classroom performance in mathematics and/or reading was then supplemented by diagnostic testing, followed by an evaluation by special educators. The information obtained by the classroom teachers, the diagnostic tests, and the special educators revealed that the level of skill in mathematics and/or reading of the present children was considerably below that expected in Grades 3 or 4. It was therefore decided that the children should receive special instruction in mathematics, reading, or both.

### General Testing Procedure

This research was part of an ongoing longitudinal study of mathematic development in children with learning difficulties and its relation to reading development. However, four reading-related

tasks included in this project are not examined in this article (see the discussion in the section *Supporting Cognitive Functions and Measures of IQ*).

Each child undertook a total of 19 tests (including the screening tasks and IQ measures) that were administered in two separate sessions: a group test session and an individual test session. Approximately 1 to 4 weeks after the group test, the individual test session was performed. The test order was the same for all children. All instructions regarding the tasks, in both test sessions, were presented orally. Ten female experimenters performed all testing.

*Group test session.* Nine tests were administered during the group test session, but only six of the tests are reported in the present article. The tasks, in order of presentation, were the screening test of mathematics, the screening test of reading, Raven's Standard Progressive Matrices test (Raven, 1976; sets B, C, and D), the place value test, the word ability test (Järpsten & Taube, 1997), and multidigit written calculation. The group test session took approximately 120 min, divided into three 40-min sessions with 10- to 15-min breaks in between. That is, the session was the length of a regular lesson in the Swedish elementary school.

*Individual test session.* The relevant tasks in this session, in order of presentation, were as follows: arithmetic fact retrieval, visual-matrix span, calculation principles, telling time, approximate arithmetic, number matching, one-step mathematic problem solving, multistep word problem solving, and trail making. The individual test session took approximately 120 min, divided into three 40-min sessions with 10- to 15-min breaks in between.

### Screening Tests and Tests of Verbal and Nonverbal IQ

Screening tests of mathematics and reading, the word ability test (Järpsten & Taube, 1997), and Raven's Standard Progressive Matrices test (Raven, 1976; sets B, C, and D) were administered in the group test session to obtain measures of mathematic ability, reading ability, and nonverbal and verbal IQ. The mathematic and reading screening tasks were included to establish that the groups of children classified as having MD and/or RD indeed performed significantly poorer than the group of normally achieving children.

*Mathematic screening test.* The screening test was a paper-and-pencil test and consisted of three subtests. All three subtests were designed so that the test items became successively more difficult. The children were instructed that they could solve the problems in any way they wanted, and that they should not struggle and spend too much time on a single problem but instead try the next problem. Paper and pencil were allowed during performance of the task. The number of correctly solved problems was used as dependent measure. The maximum score was 32.

In the first subtest, the child was asked to solve five addition problems and five subtraction problems ( $568 + 421$ ;  $658 - 437$ ;  $56 + 47$ ;  $65 - 29$ ;  $545 + 96$ ;  $384 + 278$ ;  $824 - 488$ ;  $4203 + 5825$ ;  $8010 - 914$ ;  $11305 - 5786$ ) during 8 min. The problems were presented horizontally, because this is the primary form of presentation when starting to teach children multidigit arithmetic in Sweden. The children responded in Arabic form (e.g., 103). Half the problems involved regrouping (i.e., carrying or borrowing).

The second subtest consisted of 12 arithmetic equations presented horizontally ( $61 + \_\_\_ = 73$ ;  $\_\_\_ \times 4 = 16$ ;  $\_\_\_ \times 5 = 40$ ;  $\_\_\_ + 25 = 500$ ;  $1000 - \_\_\_ = 550$ ;  $\_\_\_ - 8 = 6$ ;  $8 \times \_\_\_$



= 24; \_\_\_\_ - 50 = 50; \_\_\_\_ - 445 = 55; 13 = 6 + \_\_\_\_; 136 = \_\_\_\_ + 27; 360 = \_\_\_\_ - 610). The task was to fill in the right number so that the equation was correct. The children were allowed 7 min to complete that task. The subtest was preceded by two practice trials before the actual testing started to ensure that the children understood the task.

In the third subtest, the child was presented with an answer and two to four numbers that had to be combined with one to three arithmetic operations (addition, subtraction, multiplication) in order to obtain the predetermined answer. For example, if the answer was 12 and the three numbers were 5, 8, 9, a correct combination would be 9 + 8 - 5. Ten problems were included in this subtest (27, 113 = 140; 11, 26 = 15; 5, 8, 9 = 12; 10, 50, 90 = 30; 11, 19, 25 = 33; 4, 16, 4 = 0; 25, 19, 11 = 5; 4, 2, 5, 9 = 9; 2, 5, 30, 60 = 100; 1, 3, 8, 25 = 0). The children were allowed 5 min to complete that task. The subtest was preceded by three practice trials before the actual testing started. Cronbach's alpha coefficients for the first, second, and third subtests have been established at .82 (*n* = 110), .81 (*n* = 186), and .74 (*n* = 148), respectively, in samples of children aged between 9 and 11 years. The reliability estimate for the third subtest was calculated on 6 out of 10 items to obtain a larger sample of children. The intercorrelations among the three subtests calculated on the sample of 186 children were significant and ranged from .54 to .65 (*p* < .01).

*Reading screening test (Malmquist, 1969).* The child had to read a short fable story (600 words in length) about a turtle and a water buffalo stealing bananas from a gardener. In the text, 20 sentences were incomplete (i.e., one word was missing and replaced by three single words within parentheses). The child was asked to read the story as quickly and accurately as possible and to select one of the three words to complete the sentence in an appropriate way. The child was allowed 4 min to work with the test. The total number of correctly completed sentences was used as dependent measure. The maximum score was 20. The reported split-half reliability after Spearman-Brown correction for this task is .94 (see Malmquist, 1977).

*Word knowledge test (Järpsten & Taube, 1997).* In this multiple-choice response test, the child had to indicate the meaning of a target word by choosing one of four options (e.g., "The word pal means the same as (a) girl, (b) boy, (c) child, (d) buddy"). Each child received a test booklet that included 24 items, and the child responded by checking one of the four options. As many of the children had poor reading skills, the experimenter read each prob-

lem and all response options to the children while they followed along in the booklet. Prior to the actual testing, two practice trials were completed. Järpsten and Taube reported a Cronbach's alpha coefficient for the word knowledge test of .87 calculated on a sample of 665 children aged between 10 and 11 years. The maximum score was 24.

*Raven's Standard Progressive Matrices test.* The test consists of a series of visual pattern designs with a piece missing. The task is to select the correct piece to complete the designs from a number of options (six to eight) displayed beneath the design. The complete test includes five sets of designs (A, B, C, D, E), and each set consists of 12 items (Raven, 1976). In the present study only sets B, C, and D were used; thus, the maximum score was 36. For set B six response options were displayed, and for sets C and D eight response options. Each child received a test booklet that included 36 test items and 2 practice items. The children responded by checking one of the six to eight options on a separate answer sheet. After the two practice items had been completed, the children completed their booklets at their own pace. The specific instructions and administration procedures followed those specified in the test manual, and the reported Cronbach's alpha coefficients for the test usually lie in the range between .90 and .97 (see Raven, Court, & Raven, 1996).

More detailed information regarding mean age; gender; number of participants; and results on Raven's Standard Progressive Matrices test, the word ability test, and screening tasks for mathematics and reading for each achievement group are displayed in Table 1.

One-way analyses of variance (ANOVAs) were computed to compare the groups on measures of nonverbal (Raven) and verbal IQ (word ability test), mathematics, and reading. Because of the number of comparisons, a significance level of *p* < .01 was employed for all statistical procedures in the study.

The results of the ANOVAs and subsequent post hoc tests (Tukey-Kramer) revealed that both groups of children with mathematic difficulties (MD only and MD-RD) performed significantly lower on the mathematic screening task, *F*(3, 178) = 20.23, *p* < .01, and that the RD-only children and the MD-RD children performed significantly lower on the reading task, *F*(3, 178) = 10.71, *p* < .01, compared to the controls. However, no significant group differences emerged on Raven's Standard Progressive Matrices and the verbal ability test (all *ps* > .01).

Table 1  
*Descriptive Information for Participants by Achievement Group*

Characteristic	MD only		MD-RD		RD only		Controls	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Mean age (in months)	125	5.97	125	7.10	128	7.41	125	7.26
Raven's Standard Progressive Matrices (raw score)	20.66	5.02	21.42	4.20	22.63	4.87	21.62	6.52
Word knowledge task (raw score)	13.78	2.93	14.00	2.68	13.63	2.99	14.28	3.16
Mathematic screening task	9.24	4.16	9.90	4.45	13.97	6.06	15.89	5.44
Reading screening task	10.24	3.42	6.82	3.12	7.80	3.18	12.13	3.46
<i>N</i> (number of boys)	41 (10)		50 (21)		30 (23)		61 (33)	

*Note.* MD only = children with mathematical difficulties only (i.e., normal reading ability); MD-RD = children with both mathematical and reading difficulties; RD only = children with reading difficulties only (i.e., normal mathematical ability).

### Experimental Tasks and Procedure

**Arithmetic fact retrieval.** The task assessed automatic retrieval of arithmetic facts. The test material consisted of 12 addition combinations ( $5 + 4$ ;  $8 + 5$ ;  $7 + 4$ ;  $9 + 5$ ;  $3 + 8$ ;  $7 + 9$ ;  $4 + 6$ ;  $3 + 9$ ;  $7 + 8$ ;  $8 + 4$ ;  $3 + 6$ ;  $6 + 5$ ), 12 subtraction combinations ( $9 - 4$ ;  $7 - 4$ ;  $9 - 5$ ;  $7 - 5$ ;  $8 - 4$ ;  $9 - 7$ ;  $6 - 3$ ;  $8 - 3$ ;  $6 - 2$ ;  $9 - 3$ ;  $9 - 6$ ;  $7 - 3$ ), and 12 multiplication combinations ( $4 \times 5$ ;  $2 \times 7$ ;  $9 \times 4$ ;  $9 \times 10$ ;  $7 \times 5$ ;  $6 \times 3$ ;  $9 \times 2$ ;  $7 \times 3$ ;  $6 \times 6$ ;  $4 \times 6$ ;  $3 \times 4$ ;  $5 \times 6$ ). The task was administered on three separate sheets of paper, one for each operation. On the three sheets the 12 number combinations were presented in a column. All children started with the addition condition, followed by the subtraction condition, and the multiplication condition. The child was instructed to provide an answer right away and was encouraged to guess if the answer was not available right away. The experimenter used a stopwatch to measure the total time it took to provide answers to the 12 number combinations and to register whether the response time was longer than 3 s for each individual problem. During performance the experimenter also continually checked the child's answers and registered each error. The total number of correctly solved problems for each condition and all three conditions together with response times within 3 s were used as dependent measures (cf. Russell & Ginsburg, 1984). In addition, the overall error rate for responses within and above 3 s was used as a dependent measure. The Cronbach's alpha coefficients for the addition, subtraction, and multiplication fact retrieval tasks calculated on the present sample were .85, .87, and .86, respectively.

**Written multidigit arithmetic calculation.** This was a paper-and-pencil test and consisted of 12 vertically presented multidigit calculation problems. The child was asked to solve five addition problems and seven subtraction problems ( $258 + 341$ ;  $45 + 67$ ;  $78 - 56$ ;  $647 - 566$ ;  $459 - 279$ ;  $545 - 378$ ;  $1295 + 5437$ ;  $3874 + 8566$ ;  $26856 + 14249$ ;  $9566 - 978$ ;  $2023 - 1455$ ;  $12405 - 7748$ ). The task was designed so that the test items became successively more difficult. All but two problems ( $258 + 341$ ;  $78 - 56$ ) involved regrouping. The child was allowed 10 min to complete the task. The maximum score was 12, and the Cronbach's alpha coefficient for the task was .88, calculated on a sample of 171 children aged 9 to 11 years.

**One-step mathematic word problems.** The child was asked to solve 14 written mathematic word problems by means of paper and pencil during 15 min (e.g., "John had 65 crowns left when he had bought a book for 36 crowns. How much did he have to start with?"). In order to impose as little linguistic demand as possible, the problems were one to three sentences long and did not include any irrelevant information. All but two problems included multidigit calculation. The task was designed so that the test items became successively more difficult. The experimenter and the child sat next to each other during the administration of the task. The experimenter read the individual problems while the child followed along on the paper. If requested by the child the experimenter reread the problem. This test administration procedure was chosen because of the poor reading skills of many of the children (i.e., MD-RD and RD only), although a redundancy effect can emerge as a result of having the child listen and read the same material simultaneously (Kalyuga, Chandler, & Sweller, 2004; Sweller, 2005). The negative impact of poor reading skills on task performance was, however, considered to be worse than the neg-

ative impact of the redundancy effect (cf. Fuchs & Fuchs, 2002; Swanson & Beebe-Frankenberger, 2004). The presentation order of the problems was the same for all children, and the children responded in writing. The maximum score was 14. The alpha coefficient for the task was .82, calculated on a sample of 147 children aged 9 to 11 years. The reliability estimate was based on the first 12 items out of 14 items to obtain a larger sample of children. None of the children in the present sample attempted to solve Items 13 and 14.

**Complex multistep mathematic word problems.** The complexity of mathematic word problems can be enhanced in a number of ways, for example by including irrelevant information and/or using indirect language (Fuchs & Fuchs, 2002; Parmar et al., 1996; Russell & Ginsburg, 1984). Another way to increase the complexity is to use problems that require multiple solution steps (Fuchs & Fuchs, 2002; Parmar et al., 1996). The present study adopted this latter approach for two reasons. First, the primary interest was to examine the children's mathematic problem-solving abilities, not their linguistic abilities; thus, using word problems that were mathematically complex appeared to be the adequate approach. Second, because two groups of children (i.e., MD-RD, RD only) had RD, it was quite possible that their reading problems might mask their "true" mathematic problem-solving abilities (the operation and problem-solving dimensions; Fuchs & Fuchs, 2002) if the problems imposed high linguistic demands by including irrelevant information.

The child was asked to solve seven written mathematic word problems that required at least two calculation steps in order to reach a correct solution (e.g., "Mark weighs 38 kg. His dad weighs 35 kg more. How much do they weigh together?"). The problems were two to four sentences long and did not include any irrelevant information. Thus, the linguistic demand was low as in the one-step problems. All problems included multidigit calculation. The same administration procedure as for the one-step mathematic word problems was used. Seven was the maximum score. The alpha coefficient for the task was .72, calculated on a sample of 166 children aged 9 to 11 years.

**Place value.** The child's understanding of the base-10 number system and place value (Jordan & Hanich, 2000) was tapped by a paper-and-pencil test that consisted of three parts. In Part 1, the child was presented with four 3-digit numbers and one 4-digit number and was asked to indicate, in writing, the value of a particular digit (e.g., "What is the value of the digit 9 in 349?" "What is the value of the digit 4 in 479?" "What is the value of the digit 8 in 986?" "What is the value of the digit 6 in 6821?" "What is the value of the digit 5 in 257?"). In Part 2, the child was asked to answer, in writing, seven questions regarding which number consists of a certain number of thousands, hundreds, tens, and ones. The seven questions were presented in the following format: Indicate in writing the number that consists of 3 hundreds, 6 tens, and 3 ones (i.e., 363). In Part 3, the child was presented with five pairs of written numbers (799999–811111; 522222–288888; 9444444–4999999; 0.299999–0.611111; 0.099999–0.100000) and was required to indicate the larger number of each pair by making a circle around the larger number (see Russell & Ginsburg, 1984). The second number in the pair was positioned beneath the first number in the pair. The maximum score was 17, and Cronbach's alpha calculated on the present sample was .78.



*Calculation principles task.* This task assessed the child's understanding of three calculation principles: the commutativity principle (i.e., that the order of addends does not affect the sum), the inversion principle (i.e., that subtraction is the inversion of addition and vice versa), and the double-plus-one principle. The task was a version of the task developed by Russell and Ginsburg (1984; see also Baroody & Wilkins, 1999) and included a total of 10 problems. Three problems assessed the commutativity principle ( $26 + 32 = 58$ , so what is  $32 + 26 = ?$ ;  $15 + 19 = 34$ , so what is  $19 + 15 = ?$ ;  $48 + 21 = 69$ , so what is  $21 + 48 = ?$ ), four problems assessed the inversion principle ( $23 + 14 = 37$ , so what is  $37 - 14 = ?$ ;  $27 + 69 = 96$ , so what is  $96 - 69 = ?$ ;  $59 - 15 = 44$ , so what is  $44 + 15 = ?$ ;  $79 - 12 = 67$ , so what is  $67 + 12 = ?$ ), and three problems assessed the double-plus-one principle ( $37 + 37 = 74$ , so what is  $37 + 38 = ?$ ;  $42 + 42 = 84$ , so what is  $42 + 43 = ?$ ;  $68 + 68 = 136$ , so what is  $68 + 69 = ?$ ). The task was designed so that the answer to the first number combination (e.g.,  $26 + 32 = 58$ ) could be used to solve the second combination ( $32 + 26 = ?$ ). The 10 problems were presented in a vertical format on three separate sheets of paper, one for each principle; the experimenter read and showed each problem to the child. To prevent the child from calculating, a maximum response time of 5 s was employed (Hanich et al., 2001). If the child did not respond within the 5-s interval, the answer was considered incorrect. The maximum score was 10, and Cronbach's alpha calculated on the present sample was .78.

*Approximate arithmetic.* This task was adapted from Dehaene et al. (1999). The test material consisted of seven addition and seven subtraction combinations, and each item was accompanied by two proposed answers (e.g.,  $72$  or  $60$ ,  $52 + 17 = ?$ ;  $47$  or  $36$ ,  $57 - 23 = ?$ ). The task was to choose the answer that was closest to the correct answer. In contrast to the procedure used by Dehaene et al., the two proposed answers were presented first, followed by the arithmetic combination (e.g., first  $72 - 60$ , and then  $52 + 17 = ?$ ). The 14 combinations were presented in a vertical format on two separate sheets of paper, one for addition and one for subtraction; all children started with the addition condition. The experimenter displayed one combination at a time to the child. The child was instructed to provide an answer right away and not to calculate an exact answer to the combination. To prevent the child from calculating, a maximum response time of 5 s was employed (Hanich et al., 2001). A stopwatch was used to register whether the response time was longer than 5 s for each individual problem. If the child did not respond within the 5-s interval the answer was considered incorrect and the child was instructed to respond quicker. Two practice trials were presented to the child before the actual test phase started to ensure that the child understood the task. The maximum score was 14, and Cronbach's alpha calculated on the present sample was .80.

*Telling time.* This task was developed by the author and consisted of four analog clocks (08:30; 10:15; 15:40; 21:55) and four digital clocks (07:35; 21:00; 13:15; 01:45) presented on a paper. The analog clocks were also supplemented with information regarding whether it was the morning (a.m.) or the evening (p.m.). When telling the time of the four analog clocks, the child had to express the time in words (e.g., quarter past ten) and in digital form (e.g., 10:15). Thus, the child could receive 0, 1, or 2 points for each analog clock. When telling the time of the four digital clocks, the child had to express the time in words and also state whether it was

the morning, forenoon, afternoon, evening, or during the night. The child could receive 0, 1, or 2 points for each one of the eight clocks. Thus, the maximum score was 16, and the child responded in writing. Cronbach's alpha coefficient for the task was .88, calculated on the present sample.

*Number-matching task.* This task was used to measure general processing speed. Each child was given a sheet of paper with 30 rows of digits. Each row consisted of seven digits, with 2 digits in each row being identical. The child's task was to cross out the two identical digits in each row as fast and accurately as possible. A practice trial of seven rows was performed before the actual task started. Performance was measured by the time taken to complete all 30 rows. Test-retest reliability of .79 was established for this task based on a sample of 156 children aged 9 to 13 years.

*Visual-matrix span task.* This task was a version of Swanson's (1992) visual working-memory task. The child was presented with a number of dots in a matrix. The task was to remember the location of the dots in the matrix. A total of 32 matrices constituted the test material. The matrices were made up of different numbers of squares. Each square was 2 cm and was drawn on a white background. The dots were black, with a diameter of 1 cm. One matrix at a time was displayed on a sheet of paper for 5 s. Then the matrix was removed, and the child was asked a process question: "Were there any dots in the first column?" After answering the process question, the child was required to draw dots in the correct squares on an identical matrix. The first matrix had nine squares ( $3 \times 3$ ) and included two dots. The matrices on the next span size had nine squares ( $3 \times 3$ ) and three dots. The third span size had 12 squares and included three dots. The complexity of the matrices increased for each new span size either by increasing the size of the matrix or increasing the number of dots. The complexity ranged from a matrix of 9 squares and two dots to a matrix of 36 squares and eight dots. Two different matrices were presented for each span size. Testing stopped when the child failed to repeat both trials at any particular span length. Thus, testing proceeded as long as the child succeeded in reproducing one of the two trials of the same span length. Visual-matrix span was measured as the most complex matrix remembered correctly, plus 0.5 points if the participant managed to replicate correctly both trials in the same span length. Andersson (in press) reported a test-retest reliability of .60 for this task.

*Trail-making task.* This classic paper-and-pencil test was used to assess the ability to switch between operations or retrieval strategies, which is assumed to be an important executive function (Baddeley, 1996; Lee, Cheung, Chan, & Chan, 2000; McLean & Hitch, 1999; Miyake et al., 2000). The task included two different test conditions, A and B. In the A condition, the material consisted of 25 encircled numbers on a sheet of paper. The task was to connect the 25 circles in numerical order as fast and accurately as possible. Each child was presented with a practice trial consisting of eight circles (1-2-3-4-5-6-7-8) and was instructed to solve it before the actual trail-making tasks commenced. In the B condition, half the circles had a number in the center (1-13) and half had a letter (A-L). The children were asked to start at number 1 and make a trail with a pencil so that each number alternated with its corresponding letter (i.e., 1-A-2-B-3-C . . . 12-L-13). Each child was presented with a practice trial consisting of eight circles (1-A-2-B-3-C-4-D) and was instructed to solve it before the actual trail-making tasks commenced. The difference in solution

time for the B condition and A condition (i.e.,  $B - A$ ) was used as dependent measure. Reported reliabilities for this task lie between .60 and .90 (Lezak, 1995).

## Results

Descriptive statistics for the overall accuracy scores for the eight mathematic tasks and the three cognitive tasks for each achievement group, and correlations among the tasks, are displayed in Tables 2 and 3.

### Overall Performance on the Mathematic and Cognitive Tasks

As a first step to test the hypotheses of the present study, ANOVAs and effect sizes ( $\eta^2$ ) were calculated for the eight mathematic tasks (overall scores) and the three cognitive tasks. According to Cohen (1988), an effect size of .09 or less is considered to be small, whereas an effect size between .10 to .24 is of medium size. An effect size of .25 or larger is considered to represent a large effect. The statistics from the first 12 ANOVAs are presented in Table 4.

**Arithmetic fact retrieval.** Two ANOVAs were calculated on the arithmetic fact retrieval task, one for overall accuracy for responses within 3 s and one for overall error rate for responses within and above 3 s. A significant and large group effect emerged for overall accuracy, and the Tukey-Kramer post hoc test revealed that both groups with MD (i.e., MD only and MD-RD) obtained significantly lower scores than the controls,  $Qs(4, 178) = 9.74; 10.95, p < .01$ , respectively; and the RD-only group,  $Qs(4, 178) = 5.82; 6.59, p < .01$ , respectively. The differences between the two MD groups, and between the RD-only group and the control group, were not significant ( $ps > .01$ ). The second ANOVA also yielded a significant group effect of medium size for overall error rate. The control group made significantly fewer errors than the MD-only group and the MD-RD group,  $Qs(4, 178) = 6.96; 7.18, p < .01$ , respectively.

**Written multidigit arithmetic calculation.** The ANOVA displayed a significant achievement group effect of medium size for

the written arithmetic calculation tasks. Post hoc testing revealed that the mean difference between the MD-only group and the controls was the only difference that reached the .01 alpha level,  $Q(4, 178) = 6.62, p < .01$ , whereas the MD-RD group only showed a tendency toward a significant group difference,  $Q(4, 178) = 4.17, p < .05$ .

**Mathematic word problems.** An ANOVA on the one-step mathematic word problem-solving task showed a significant medium size group effect. Post hoc testing revealed that both groups with MD (i.e., MD only and MD-RD) obtained significantly lower scores than the controls,  $Qs(4, 178) = 8.60; 6.09, p < .01$ , respectively; and the RD-only group,  $Qs(4, 178) = 6.75; 4.53, p < .01$ , respectively. The two MD groups did not differ from each other, and neither did the control group and the RD-only group ( $ps > .01$ ). A similar pattern of performance emerged on the multistep mathematic word problem-solving task; that is, both the MD-only group and the MD-RD group performed worse than the controls,  $Qs(4, 178) = 8.53; 8.18, p < .01$ , respectively; and the RD-only group,  $Qs(4, 178) = 6.16; 5.71, p < .01$ , respectively. The children in the RD-only group performed on a par with the control children ( $ps > .01$ ).

**Place value.** A small but significant group effect was obtained on the place value task; however, the Tukey-Kramer post hoc test revealed that none of the group differences reached the .01 alpha level.

**Calculation principles.** On the calculation principles task, a medium-sized significant group effect emerged. Subsequent post hoc testing showed that the controls outperformed both children with MD only and children with MD-RD,  $Qs(4, 178) = 4.88; 6.34, p < .01$ , respectively. In addition, the RD-only group obtained significantly higher scores on this task than the MD-RD group,  $Q(4, 178) = 5.26, p < .01$ . No other group differences reached significance on this task ( $ps > .01$ ).

**Approximate arithmetic.** The ANOVA yielded a significant group effect of medium size on the approximate arithmetic task. Both the MD-only group and the MD-RD group performed significantly worse than the controls,  $Qs(4, 178) = 6.13; 7.23, p < .01$ , respectively. The MD-RD group was also outperformed by

Table 2  
Mean Performance and Standard Deviations on the Mathematic Tasks by Achievement Group

Task	MD only		MD-RD		RD only		Controls	
	M	SD	M	SD	M	SD	M	SD
Arithmetic fact retrieval, solved problems within 3 s	16.34	8.40	15.68	8.20	23.93	7.62	27.02	6.68
Total number of errors	6.27	5.00	6.16	5.72	3.33	4.07	2.00	1.80
Written arithmetic calculation	3.24	2.29	4.38	2.94	5.10	2.96	6.05	3.36
Approximate arithmetic	5.05	2.93	4.72	3.37	7.30	3.24	7.93	3.48
Place value	12.98	2.74	12.88	2.78	14.10	1.67	14.10	1.60
Calculation principles	6.07	2.42	5.70	2.56	7.73	2.16	7.72	2.24
One-step mathematic word problems	4.12	2.57	5.24	2.72	7.27	3.36	7.49	2.53
Multistep mathematic word problems	0.95	1.20	1.12	1.27	2.50	1.83	2.75	1.61
Telling time	8.76	4.70	8.06	3.84	11.83	4.13	12.54	3.08
Visual-matrix span	5.06	2.04	5.35	1.67	5.58	2.33	6.12	2.12
Trail making	88.95	50.22	96.22	48.49	80.53	41.06	70.85	47.82
Visual number matching	154.12	42.30	164.08	48.52	140.87	40.68	127.67	33.61

Note. MD only = children with mathematical difficulties only (i.e., normal reading ability); MD-RD = children with both mathematical and reading difficulties; RD only = children with reading difficulties only (i.e., normal mathematical ability).



Table 3  
Correlations Among the Tasks Used in the Study

Task	1	2	3	4	5	6	7	8	9	10	11	12
1. Arithmetic fact retrieval	—	.46	.55	.26	.40	.61	.56	.47	.12	-.17	-.36	.22
2. Written calculation		—	.22	.22	.24	.61	.47	.38	.26	-.22	-.36	.34
3. Approximate arithmetic			—	.22	.37	.31	.35	.32	.06	-.20	-.26	.21
4. Place value				—	.20	.43	.27	.37	.17	-.15	-.17	.23
5. Calculation principles					—	.36	.38	.48	.12	-.13	-.23	.18
6. One-step word problems						—	.70	.60	.22	-.25	-.32	.35
7. Multistep word problems							—	.53	.24	-.24	-.27	.30
8. Telling time								—	.24	-.26	-.34	.25
9. Visual-matrix span									—	-.24	-.29	.25
10. Trail making										—	.30	-.06
11. Visual number matching											—	-.22
12. Age												—

Note.  $n = 182$ ,  $df = 180$ . Correlation coefficients larger than .19 were significant at the 1% level.

the RD-only group,  $Q(4, 178) = 4.80$ ,  $p < .01$ . No other group differences were significant on this task ( $ps > .01$ ).

*Telling time.* Finally, a significant group effect of medium size was found on the time-telling task. Both groups of children with MD (MD only and MD-RD) performed significantly lower compared to the controls,  $Qs(4, 178) = 6.16$ ;  $5.71$ ,  $p < .01$ , respectively; and the RD-only group,  $Qs(4, 178) = 6.16$ ;  $5.71$ ,  $p < .01$ , respectively. No significant group differences emerged, however, between the controls and the RD-only group ( $ps > .01$ ).

*Supporting cognitive functions.* In order to examine expected group differences in basic cognitive functions ANOVAs were calculated on the three cognitive tasks employed in the study. A significant group effect of medium size was obtained on the number-matching task, and both the MD-only group and the MD-RD group performed significantly worse than the controls,  $Qs(4, 178) = 4.49$ ;  $6.54$ ,  $p < .01$ , respectively. No significant group differences emerged on the visual-matrix span task and the trail-making task ( $ps > .01$ ).

Additional Analyses

In order to obtain a more detailed picture of the MD displayed by the MD-only children and the MD-RD children, additional

analyses were performed on the arithmetic fact retrieval task, the place value task, the calculation principles task, the approximate arithmetic task, and the time-telling task.

*Arithmetic fact retrieval and the influence of general processing speed.* The significantly slower performance of the MD-only and MD-RD children on the visual number-matching task indicated that they had a weakness in general processing speed. To examine the possibility that these children's poor performance in arithmetic fact retrieval was a consequence of general slowness, an analysis of covariance (ANCOVA) with the number-matching task as a covariate was performed on the arithmetic fact retrieval task. The assumption regarding homogeneous regression slopes between groups for the covariate was fulfilled ( $p = .40$ ). The ANCOVA revealed a significant group effect,  $F(3, 177) = 19.45$ ,  $p < .01$ ,  $\eta^2 = .25$ ,  $MSE = 55.94$ , and significant contributions also emerged for the number matching task,  $F(1, 177) = 10.24$ ,  $p < .01$ ,  $\eta^2 = .06$ ,  $MSE = 55.94$ . Post hoc testing showed that the performance of the MD-only group (adjusted  $M = 16.70$ ) and the MD-RD group (adjusted  $M = 16.48$ ) was significantly lower than that of the controls (adjusted  $M = 26.23$ ),  $Qs(4, 177) = 8.73$ ;  $9.46$ ,  $p < .01$ , respectively; and the RD-only group (adjusted  $M = 23.72$ ),  $Qs(4, 177) = 5.40$ ;  $5.80$ ,  $p < .01$ , respectively.

Table 4  
Results of the Analysis of Variance Performed on the Mathematic and Cognitive Tasks

Task	$F(3, 178)$	MSE	$\eta^2$	Post hoc
Arithmetic fact retrieval, solved problems within 3 s	27.02**	58.85	.31	CON = RD only > MD only = MD-RD
Total number of errors	12.35**	18.42	.17	CON > MD only = MD-RD, CON = RD only
Written arithmetic calculation	7.81**	8.79	.12	CON > MD only = MD-RD, CON = RD only
One-step word problems	15.79**	7.52	.21	CON = RD only > MD only = MD-RD
Multistep word problems	18.60**	2.19	.24	CON = RD only > MD only = MD-RD
Place value	4.10**	5.14	.06	CON = RD only = MD only = MD-RD
Calculation principles	9.54**	5.59	.14	CON > MD only = MD-RD, RD only > MD-RD
Approximate arithmetic	11.73**	10.54	.16	CON > MD only = MD-RD, RD only > MD-RD
Telling time	16.13**	14.99	.21	CON = RD only > MD only = MD-RD
Number matching	7.90**	1,700.56	.12	CON > MD only = MD-RD
Visual-matrix span	2.57	4.10	.04	CON = RD only = MD only = MD-RD
Trail making	2.85	2,259.38	.05	CON = RD only = MD only = MD-RD

Note. CON = controls; RD only = children with reading difficulties only (i.e., normal mathematical ability); MD only = children with mathematical difficulties only (i.e., normal reading ability); MD-RD = children with both mathematical and reading difficulties  
\*\*  $p < .01$ .

A 3 (type of number combination)  $\times$  4 (achievement group) mixed ANOVA was also performed on measures of arithmetic fact retrieval to examine possible interaction effects. However, no significant interaction effect or main effect of type of number combination (i.e., addition, subtraction, multiplication) was found ( $p < .01$ ).

*Place value, calculation principles, and approximate arithmetic.* For the same reason as for the arithmetic fact retrieval task, 3  $\times$  4 mixed ANOVAs were performed on the three submeasures of the place value task and the three submeasures of the calculation principles task. However, because the submeasures on these two tasks included different numbers of items, the ANOVAs were performed on proportion correct instead of raw scores. A 2 (addition vs. subtraction)  $\times$  4 (achievement group) mixed ANOVA was also performed on the approximate arithmetic task. The results of the three mixed ANOVAs did not display any significant interactions; that is, the group mean differences did not vary between the different submeasures of the three tasks.

*Telling time.* A fifth 2 (analog clock vs. digital clock)  $\times$  4 (achievement group) mixed ANOVA was computed on the time-telling task. This analysis showed a significant interaction between type of clock and group,  $F(3, 178) = 3.89, p < .01, \eta^2 = .06, MSE = 1.82$ . Descriptive statistics for the two submeasures are presented in Table 5. Test of simple effects revealed significant group effects for both the analog clocks,  $F(3, 356) = 19.82, p < .01, MSE = 2.21$ ; and the digital clocks,  $F(3, 356) = 36.31, p < .01, MSE = 2.21$ . For the analog clocks, subsequent Tukey-Kramer tests showed that the MD-RD group performed worse than the RD-only group,  $Q(4, 356) = 5.58, p < .01$ ; as well as the control group,  $Q(4, 356) = 10.25, p < .01$ ; whereas the MD-only group performed worse than the controls,  $Q(4, 356) = 7.28, p < .01$ ; but not the RD-only group ( $p > .01$ ). On the digital clocks, in contrast, both the MD-RD group and the MD-only group performed worse than the controls,  $Qs(4, 356) = 11.13; 11.02, p < .01$ , respectively; and the RD-only group,  $Qs(4, 356) = 9.43; 9.48, p < .01$ , respectively. Furthermore, test of simple effect showed that the RD-only group obtained significantly higher scores on the digital clocks compared to the analog clocks,  $F(1, 356) = 8.73, p < .01, MSE = 2.21$ , whereas the three other groups performed equally on the two types of clocks ( $p > .01$ ).

### *Written Multidigit Calculation and the Influence of Arithmetic Facts Retrieval*

Proficiency in written multidigit calculation is dependent on the ability to quickly retrieve arithmetic facts from long-term memory

Table 5  
Mean Performance and Standard Deviations on the Time-Telling Task by Achievement Group

Type of clock	MD only		MD-RD		RD only		Controls	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Analog	4.49	2.06	3.98	1.98	5.33	2.26	6.03	1.63
Digital	4.07	2.72	4.18	2.39	6.47	2.26	6.41	1.90

*Note.* MD only = children with mathematical difficulties only (i.e., normal reading ability); MD-RD = children with both mathematical and reading difficulties; RD only = children with reading difficulties only (i.e., normal mathematical ability).

(Ashcraft, 1992, 1995; Geary, 1993; Jordan, Hanich, & Uberti, 2003; McCloskey, Caramazza, & Basili, 1985). This connection is demonstrated by the correlation of .46 found between arithmetic fact retrieval and multidigit calculation (see Table 3). Therefore, in order to investigate whether the poor performance of the MD-only group and MD-RD group on the written multidigit calculation was due to their poor skill in arithmetic fact retrieval, an ANCOVA was computed with the arithmetic fact retrieval task as a covariate. The assumption regarding homogeneous regression slopes between groups for the covariate was fulfilled ( $p = .44$ ). The ANCOVA did not yield a significant group effect ( $p > .01$ ); the influence of the covariate, however, was significant,  $F(1, 177) = 24.38, p < .01, \eta^2 = .12, MSE = 7.77$ . Thus, the group differences in written multidigit calculation were completely accounted for by group differences in arithmetic fact retrieval.

### *Word Problem Solving and the Influence of Calculation Ability and Understanding Calculation Principles*

Mathematic word problem solving involves a complex interplay of language comprehension, problem representation, selection of calculation operations (i.e., understanding of calculation principles), and execution of calculation operations (Kintsch & Greeno, 1985; Mayer & Hegarty, 1996; Swanson, 2004). As such, it is possible that the observed problems that the MD-only group and the MD-RD group displayed on the two problem-solving tasks were a consequence of problems with some of the previously mentioned processes. In an attempt to examine this possibility, two ANCOVAs were performed with the written multidigit calculation task and the calculation principles task as covariates. The assumptions regarding homogeneous regression slopes between groups was met for both mathematic word problem tasks ( $p > .23$ ). The ANCOVA on the simple one-step mathematic word problem-solving task demonstrated a significant group effect,  $F(3, 176) = 5.45, p < .01, \eta^2 = .08, MSE = 5.05$ ; and significant contributions also emerged for the calculation task,  $F(1, 176) = 72.33, p < .01, \eta^2 = .29, MSE = 5.05$ ; and the calculation principles task,  $F(1, 176) = 7.63, p < .01, \eta^2 = .04, MSE = 5.05$ . Post hoc testing showed that the performance of the MD-only group (adjusted  $M = 5.03$ ) was significantly lower than that of the controls (adjusted  $M = 6.70$ ),  $Q(4, 176) = 4.72, p < .01$ ; and the RD-only group (adjusted  $M = 6.94$ ),  $Q(4, 176) = 4.53, p < .05$ , when the .05 alpha level was used. The performance level of the MD-RD children was not significantly lower than that of the controls or the RD-only group ( $p > .01$ ).

A significant group effect emerged on the multistep mathematic word problem-solving task,  $F(3, 176) = 7.82, p < .01, \eta^2 = .12, MSE = 1.81$ . Moreover, the calculation task,  $F(1, 176) = 26.28, p < .01, \eta^2 = .13, MSE = 1.81$ , and the calculation principles task,  $F(1, 176) = 8.22, p < .01, \eta^2 = .04, MSE = 1.81$ , turned out to be significant covariates. Follow-up comparisons revealed that the MD-RD group (adjusted  $M = 1.32$ ) and MD-only group (adjusted  $M = 1.33$ ) scored significantly poorer than the control group (adjusted  $M = 2.42$ ),  $Qs(4, 176) = 4.47$  and  $4.54, p < .01$ , respectively.

As both word problem-solving tasks displayed a rather strong correlation ( $rs = .61; .56$ ; see Table 3) with the arithmetic fact retrieval task and the written multidigit calculation task, two ANCOVAs were calculated to examine whether the observed



group differences in word problem solving could be accounted for by skill in arithmetic fact retrieval and multidigit calculation skill. The regression slopes were homogeneous for both covariates and both word problem-solving tasks ( $p > .16$ ).

On the one-step word problems no significant group effect emerged ( $p > .01$ ). The influence of the arithmetic fact retrieval task,  $F(1, 176) = 28.56, p < .01, \eta^2 = .14, MSE = 4.53$ ; and the written multidigit calculation task,  $F(1, 176) = 50.14, p < .01, \eta^2 = .22, MSE = 4.53$ , was, however, significant.

A significant group effect was found on the multistep word problem-solving task,  $F(3, 176) = 4.02, p < .01, \eta^2 = .06, MSE = 1.73$ ; and significant contributions also emerged for the arithmetic fact retrieval task,  $F(1, 176) = 16.35, p < .01, \eta^2 = .09, MSE = 1.73$ ; and the multidigit calculation task,  $F(1, 176) = 15.91, p < .01, \eta^2 = .08, MSE = 1.73$ . Post hoc testing showed that the performance of the MD-only group (adjusted  $M = 1.43$ ) and the MD-RD group (adjusted  $M = 1.47$ ) was significantly lower than that of the controls (adjusted  $M = 2.24$ ),  $Qs(4, 176) = 3.96; 3.97$ , respectively, when the .05 alpha level was used.

*Further Analyses of Multistep Word Problem Solving With Hierarchical Regression Analysis*

To further examine factors that might account for ability group differences in multistep word problem solving, two hierarchical regression analyses were calculated. That is, could the observed group differences in multistep word problem solving be accounted for by children's problems with fact retrieval, multidigit calculation, and understanding calculation principles? Two group contrast variables were created, one that captured the difference between the MD-only group and the control group (MD only vs. CON) and one that captured the difference between the MD-RD group and the control group (MD-RD vs. CON). These two contrast variables were coded as a trichotomy: +1, -1, 0 (Pedhazur, 1982). In the first contrast variable, MD only versus CON, the MD-only group was coded as +1, the control group as -1, and the RD-only group and the MD-RD group as 0. In the second contrast variable, the MD-RD group was coded as +1, the controls as -1, and the remaining two groups as 0. In the first regression model, the first block included the arithmetic fact task, the calculation task, and the calculation principles task; these three tasks were entered simultaneously. The second block included the MD only versus CON contrast variable, and the third block included the MD-RD versus CON contrast variable. The first block in the second model included the same tasks as the first model, whereas the order of the contrast variables in Blocks 2 and 3 was inverted compared to Model 1. The results of the analysis are presented in Table 6.

The three variables included in the first block accounted together for 39% of the variance in mathematic word problem performance. More important, both contrast variables, MD only versus CON, and MD-RD versus CON, accounted for 2% of the variance when entered into the regression model after the arithmetic fact retrieval task, the calculation task, and the calculation principles task, when the .05 alpha level was used (i.e.,  $ps = .023; .033$ , respectively).

To summarize, compared to the controls, and in many cases also the RD-only group, children with MD only or MD-RD performed significantly poorer on a number of tasks tapping different domains of mathematics, such as arithmetic fact retrieval, approxi-

Table 6  
*Hierarchical Regression Analysis of Performance on the Multistep Mathematic Word Problem-Solving Task*

Model	Predictor	dfs	$\Delta R^2$	F	$R^2$
Model 1					
Block 1	Calculation principles	3,170	.39**	35.68**	.39**
	Arithmetic fact retrieval				
	Multidigit calculation				
Block 2	MD only vs. CON	4,169	.02*	28.75**	.41**
Block 3	MD-RD vs. CON	5,168	.00	23.41**	.41**
Model 2					
Block 1	Calculation principles	3,170	.39**	35.68**	.39**
	Arithmetic fact retrieval				
	Multidigit calculation				
Block 2	MD-RD vs. CON	4,169	.02*	28.49**	.41**
Block 3	MD only vs. CON	5,168	.00	23.41**	.41**

Note. MD only = children with mathematical difficulties only (i.e., normal reading ability); CON = controls; MD-RD = children with both mathematical and reading difficulties.  
\*  $p < .05$ . \*\*  $p < .01$ .

mate arithmetic, calculation principles, mathematic word problem solving, and telling time. The MD-only and MD-RD children's lower performance on the arithmetic fact retrieval task was not removed when general processing speed was controlled for in an ANCOVA. The MD-only children also obtained lower scores on the written multidigit calculation task in comparison to the controls, but this group difference was eliminated when the arithmetic fact retrieval task was used as a covariate in the ANCOVA. The group differences on the one-step mathematic word problem-solving task was eliminated when the ANCOVA included the multidigit calculation task and the calculation principles task, or the multidigit calculation task in combination with the arithmetic fact retrieval task, as covariates. In contrast, both groups of children with MD performed significantly poorer than the controls on the multistep mathematic word problem-solving task even when the influence of the calculation principles task, the arithmetic fact retrieval task, and the multidigit calculation task were controlled for in the regression model. No significant group effects emerged for the place value task. Furthermore, the two groups with MD were significantly slower when performing the number-matching task but performed on a par with the controls on the visual-matrix span task and the trail-making task.

Discussion and Conclusions

The current study was an attempt to extend previous findings regarding the mathematic competencies of children with MD only, MD-RD, and RD only. As predicted, the children with MD only and MD-RD performed significantly poorer than the normally achieving children on all mathematic tasks, with one exception: the place value task. The results are now discussed in relation to the different domains of mathematics.

*Arithmetic Fact Retrieval*

The present results corroborate previous findings that a main problem for children with MD, whether they have reading problems or not, is to solve simple arithmetic fact problems (e.g., 5 +

4;  $9 - 4$ ;  $3 \times 4$ ) by direct retrieval of the answer from long-term memory (Andersson & Lyxell, 2007; Geary et al., 1991; Jordan & Montani, 1997; Ostad, 1997; Siegler, 1988; Temple, 1991).

An important finding is that both groups of children with MD still performed significantly poorer than the controls when the effect of general processing speed was controlled for in the analysis. Controlling for processing speed only reduced effect sizes with 6% from .31 to .25. Thus, the large effect size obtained for this task ( $\eta^2 = .31$ ) indicates that children with MD or MD-RD have a severe deficit related to their ability to establish and quickly retrieve arithmetic facts from long-term memory that cannot be attributed to general slowness. Not only did the two MD groups solve considerably fewer simple arithmetic fact problems by direct retrieval of the answer, they also made significantly more errors than the control group, even when the answer was not provided within the 3-s response interval. This latter finding indicates either that these children have incorrect answers associated with the problems stored in long-term memory, or that they have procedural problems preventing them from calculating correct answers by some other strategy (e.g., verbal counting).

### *Written Multidigit Calculation*

As expected, the MD-only group displayed a weakness in solving vertically presented multidigit calculation problems by means of paper and pencil (e.g.,  $258 + 341$ ;  $545 - 378$ ). The finding that the MD-RD group only showed a tendency ( $p < .05$ ) toward lower scores on the multidigit calculation task in comparison to the controls was, however, unexpected. Nevertheless, an important finding is that the lower performance of the MD-only group (and the MD-RD group) was eliminated when the influence of arithmetic fact retrieval was controlled for in the analysis. Thus, the medium-sized group differences ( $\eta^2 = .12$ ) in written multidigit calculation were completely accounted for by group differences in arithmetic fact retrieval ( $\eta^2 = .12$ ), suggesting that deficits in forced arithmetic fact retrieval are one underlying factor to the problems children with MD display with multidigit calculation. This interpretation is consistent with models of number processing and calculation (McCloskey, 1992; McCloskey et al., 1985) and research literature concerning the relationship between fact retrieval and multidigit calculation (Ashcraft, 1992, 1995; Geary, 1993; Gersten, Jordan, & Flojo, 2005; Jordan, Hanich, & Uberti, 2003).

### *Mathematic Word Problem Solving*

Mathematic word problem solving skill was examined with two types of written problems: one-step problems (e.g., "John had 65 crowns left when he had bought a book for 36 crowns. How much did he have to start with?") and multistep problems (e.g., "Mark weighs 38 kg. His dad weighs 35 kg more. How much do they weigh together?"). The difference between the two tasks was that the latter required at least two calculation steps in order to reach a correct solution, whereas the former could be solved in one calculation step.

Consistent with the prediction and previous research, the present study provides further empirical evidence that mathematic problem solving is an area of severe weakness for both groups of children with MD (cf. Fuchs & Fuchs, 2002; Jordan & Hanich,

2000; Parmar et al., 1996; Russell & Ginsburg, 1984), as they performed significantly worse than the control group and the RD-only group on both problem-solving tasks (one-step and multistep problems). It should be noted that the effect size measures for the two word problem-solving tasks were .21 and .24, respectively, which are close to large effect sizes (i.e.,  $\eta^2 = .25$ ). An important and novel finding is that the MD-only group still had a significantly lower level of performance when the influence of written multidigit calculation ( $\eta^2 = .29$ ) and understanding of calculation principles ( $\eta^2 = .04$ ) were controlled for in the analysis, whereas the MD-RD group now performed on a par with the controls. A small group effect size was now obtained ( $\eta^2 = .08$ ). The difference between the MD-only group and the controls on the one-step word problem-solving task disappeared when the written multidigit calculation task ( $\eta^2 = .22$ ) in combination with the arithmetic fact retrieval task ( $\eta^2 = .14$ ) were included in the ANCOVA as covariates.

Another key finding is that both groups of MD continued to perform significantly poorer than the controls on the multistep mathematic word problem-solving task when the effect of multidigit calculation skill ( $\eta^2 = .13$ ) and understanding of calculation principles ( $\eta^2 = .04$ ) were controlled for in the analysis. The effect size for ability group was now reduced to .12. Both groups of MD still performed poorer on the multistep word problem-solving task ( $\eta^2 = .06$ ) when the influence of the arithmetic fact retrieval task ( $\eta^2 = .09$ ), and the multidigit calculation task ( $\eta^2 = .08$ ), was controlled. Furthermore, the two hierarchical regression models revealed a small ( $\Delta R^2 = .02$ ) but significant difference between the controls and the two groups with MD, even when the influence of the calculation principles task, the arithmetic fact retrieval task, and the multidigit calculation task were controlled for in the regression model.

The results related to the one-step mathematic problem-solving task suggest that the MD-only and the MD-RD children's problem with this type of mathematic task is completely accounted for by their problems with multidigit calculation, arithmetic fact retrieval, and understanding calculation principles.

The fact that the performance of the MD-only and the MD-RD children on the multistep problem-solving task continued to be significantly lower than the controls' performance after controlling for arithmetic fact retrieval, calculation skill, and understanding of calculation principles indicates that the poor problem-solving skills of children with MD to some extent also are related to other processes involved in mathematic problem solving. For example, the integration of the different propositions (i.e., relation, number, and question propositions) contained in the problem into a problem representation, and developing a solution plan, are two fundamental processes of mathematic problem solving that might be impaired in children with MD (Kintsch & Greeno, 1985; Mayer & Hegarty, 1996; Swanson, 2004).

Contrary to the prediction, the MD-only group did not outperform the MD-RD group on the two problem-solving tasks. One plausible account of this negative finding is that the problem-solving tasks used in the present study were more complex in terms of calculation operations and problem-solving requirements compared to those in previous studies, whereas the linguistic demands were low (e.g., Fuchs & Fuchs, 2002; Jordan & Hanich, 2000; Jordan, Hanich, & Kaplan, 2003). Hence, the combination of low linguistic demands and high operational and problem-solving



demands seems to eliminate the linguistic advantage the MD-only children have been found to have over the MD-RD children. Thus, the major obstacle for the MD-RD children in solving the present word problem-solving tasks is not related to linguistic demands.

### *Place Value*

The child's understanding of the base-10 number system and place value was tapped by three subtests requiring the child to identify the value of a particular digit in a multidigit number (e.g., "What is the value of the digit 9 in 349?"); to identify which number consists of a certain number of thousands, hundreds, tens, and ones; and to identify the larger number of pairs of numbers (e.g., 799999–811111). In contrast to studies by Jordan and colleagues (Hanich et al. 2001; Jordan & Hanich, 2000; Jordan, Hanich, & Kaplan, 2003), the present study found that the two MD groups performed on a par with the control group on the place value task. This finding is consistent with results reported by Russell and Ginsburg (1984), who found that fourth graders with MD-RD did not show a deficit in the understanding of place value. Thus, the present results, in combination with previous studies on second, third, and fourth graders with MD, suggest that children with MD catch up with their normally performing peers on this fundamental aspect of mathematics as they grow older.

### *Calculation Principles*

To demonstrate his or her understanding of the relationships within and between arithmetic operations, the child had to solve problems tapping three different calculation principles: the commutativity principle ( $26 + 32 = 58$ , so what is  $32 + 26 = ?$ ), the inversion principle ( $23 + 14 = 37$ , so what is  $37 - 14 = ?$ ), and the double-plus-one principle ( $37 + 37 = 74$ , so what is  $37 + 38 = ?$ ). The present data showed that children in Grades 3 and 4 with MD only and MD-RD have a less developed understanding of the relationships within and between arithmetic operations (i.e., calculation principles). The observed group effect was of medium size, as the effect size amounted to .14. These findings are consistent with results reported by Jordan and colleagues (Hanich et al., 2001; Jordan, Hanich, & Kaplan, 2003) but contradict the findings of Russell and Ginsburg (1994). One plausible account of this inconsistency is that the children studied by Russell and Ginsburg were older (fourth graders) than the children in the present study (third and fourth graders) and in the studies reported by Jordan and colleagues (second and third graders). Thus, it is possible that most children with MD, as they get older, obtain an understanding of calculation principles comparable to that of normally achieving children; that is, that they grow out of their earlier achievement weaknesses in this specific domain of mathematics. To test this explanation a two-way ANOVA with ability group and grade level as factors was calculated to display a possible interaction effect. However, the ANOVA did not provide support for the account, as the interaction between ability group and grade was nonsignificant ( $p > .01$ ). Thus, the group mean differences did not vary as a function of grade.

### *Approximate Arithmetic*

This domain of mathematics was examined by asking the child to choose the closest approximate answer, out of two proposed

answers, to two-digit addition and subtraction combinations within 5 s (e.g., Which answer of 72 or 60 is closest to the correct answer to  $52 + 17 = ?$ ). Although the obtained effect size for this task "only" amounted to .16, the present results provide additional support to the claim that problems with estimation might be a core deficit for children with MD only and MD-RD (cf. Hanich et al., 2001; Jordan, Hanich, & Kaplan, 2003). As the performance of approximate arithmetic is assumed to require a visual-spatial representation in the form of a mental number line (Dehaene et al., 1999), future research should, therefore, test the hypothesis that children with MD have a deficit related to visual-spatial representation, and should test whether this deficit is general in nature or specifically related to numerical magnitude (cf. Jordan, Hanich, & Kaplan, 2003). Results in line with this hypothesis have been demonstrated by Andersson and Lyxell (2007), who found that children with MD only and MD-RD have poor visual working memory compared to age-matched controls. The MD-RD children were also shown to have poorer visual working memory than controls 1 year younger in age.

### *Telling Time*

Time-telling skills were assessed by requiring the child to express the time of four analog clocks in words (e.g., quarter past ten) and in digital form (e.g., 10:15) and to express the time in words of four digital clocks.

A new and important finding is that children with MD only and MD-RD have rather substantial problems with telling time ( $\eta^2 = .21$ ). The time-telling problems were equal for analog and digital clocks. According to previous studies, most children are able to identify almost all analog and digital times (e.g., 21:00; 08:30; 10:15; 15:40; 21:55; 07:35) at the age of 8 to 10 years (Case et al., 1986; Friedman & Laycock, 1989; Siegler & McGilly, 1989; Vakali, 1991). Thus, one would expect that the children who participated in the present study should manage to correctly identify most of the clocks used in the study. The performance of the controls and the RD-only children were consistent with these previous studies. The MD-only and MD-RD children, in contrast, only managed to obtain approximately 50% of the maximum score (8 out of 16). Of particular interest is their low performance for the digital clocks, as previous research has demonstrated that the ability to tell the time of digital clocks develops much earlier than that for analog clocks (7 to 8 years; Friedman & Laycock, 1989).

The time-telling problems displayed by both groups of MD children are logical considering that telling time is a complex skill that involves a number of operations that have much in common with basic arithmetic (Friedman & Laycock, 1989; Siegler & McGilly, 1989). As such, deficit in time telling ought to be found in children with MD. Because the MD-only and MD-RD children displayed a comparable number of problems with the analog and digital clocks, which are assumed to involve different processes, this suggests that they have deficits related to establishing and retrieving memory representations (analog and digital clocks; Friedman & Laycock, 1989; Siegler & McGilly, 1989) as well as counting procedures (analog clocks; Friedman & Laycock, 1989). These two types of deficit are also clearly demonstrated within the other domains of mathematics (e.g., fact retrieval, multidigit calculation, word problem solving).



### *Basic Cognitive Functions*

In addition to the different mathematic domains, the present study examined three basic cognitive functions: visual working memory, general processing speed, and executive functions. Visual working memory was tapped by a dual task (visual-matrix span) requiring the child to remember the location of dots in a matrix and to answer a process question. Visual working-memory capacity was measured as the most complex matrix remembered correctly. A number-matching task that involved crossing out two identical digits in rows of seven digits as fast and accurately as possible was used to measure general processing speed. The classic trail-making task was used to assess one important executive function, the ability to switch between operations or retrieval strategies. The task was to connect as fast and accurately as possible 25 circles, 13 with a number in the center and 12 with a letter in the center, so that each number alternated with its corresponding letter (i.e., 1-A-2-B-3-C . . . 12-L-13).

Similar to previous studies, this study demonstrates that the two groups with MD have deficiencies connected to general processing speed, as they performed the number-matching task significantly slower than the controls ( $\eta^2 = .12$ ; cf. Andersson & Lyxell, 2007; Bull & Johnston, 1997; Swanson & Beebe-Frankenberger, 2004). In contrast to Andersson and Lyxell and Swanson (1994), the two MD groups performed on a par with the control group on the visual-matrix span task. This finding, however, is consistent with results reported by Swanson and Beebe-Frankenberger and by Swanson and Sachse-Lee (2001). The data concerning the trail-making task do not replicate findings reported by Andersson and Lyxell or McLean and Hitch (1999), suggesting that children with MD only and MD-RD have a problem related to the executive function of switching between operations or retrieval strategies.

The combined results of the three cognitive tasks suggest that the MD displayed by the MD-only and MD-RD children cannot be attributed to deficits in visual working memory and the switching aspect of executive function. Neither are the processing speed problems a plausible underlying factor to these children's difficulties with mathematics, as the effect size was only .12, and controlling for processing speed did not eliminate the large group effect on the arithmetic fact retrieval task but only reduced it by 6%. Furthermore, the correlations between the different mathematic tasks and the number-matching task were rather low, ranging in size from  $-.17$  to  $.36$  (see Table 3).

### *Children With RD Only and Mathematics*

The RD-only children performed at the same level as the control group on all tasks of mathematics used in the study. Thus, in line with the majority of previous research, the present study does not provide any data suggesting that children with RD only have any weakness in mathematics, not even forced fact retrieval or word problem solving (Jordan & Hanich, 2000; Jordan, Hanich, & Kaplan, 2003). However, the possibility still remains that this group of children encounters problems with areas of mathematics that are mediated via language (i.e., reading) in later elementary school (see Jordan, Hanich, & Kaplan, 2003).

### *Educational Implications*

The present findings have at least two important educational implications. First, as skill in arithmetic fact retrieval seems to

have a major impact on more complex mathematic skills (e.g., multidigit calculation, word problem solving), and deficits in fact retrieval are a cardinal characteristic of children with MD, this suggests that mathematic instructions should place more emphasis on teaching children with MD this basic skill. In order to promote word problem-solving skill in children with MD, special instructions should also emphasize multidigit calculation fluency and understanding of calculation principles. The present findings indicate that, in addition to these more basic skills, it is important to provide direct instructions concerning conceptual aspects of problem solving, such as problem representation and developing solution plans. Thus, in order to help children with MD to overcome their problems with mathematic word problem solving, the remedial instructions should focus on both basic and complex mathematic skills that facilitate word problem solving (Geary, 2004; Jordan, Hanich, & Uberti, 2003). Second, the ability to tell the time is an important area of mathematics that ought to get more attention when providing special instructions for children with MD. This skill shares many features with basic arithmetic, a statement that is supported by the present correlations between the time-telling task and the other mathematic tasks (see Table 3; cf. Friedman & Laycock, 1989; Siegler & McGilly, 1989). Hence, besides the value of mastering this skill that is used on a daily basis, improving the time-telling skills of MD children might have positive effects on other areas of mathematics.

### *Limitations of the Present Study and Further Research*

At least three limitations of the present study should be acknowledged. One limitation is that the MD-only group only included 10 boys, which may reduce the possibility of generalizing the present findings to boys with MD only. Another potential limitation of the present study is its use of time limits for the multidigit calculation task and the mathematic word problem-solving tasks. Although the present findings indicate that MD in children cannot be attributed to general slowness, and the tasks in question were designed so that the test items became successively more difficult, there is always a possibility that the MD-only and MD-RD children might have obtained higher scores if they had had the opportunity to attempt all test items. A third limitation is the lack of verbal working-memory tasks. The visual working-memory capacity of the MD children was shown to be equal to that of the controls; it is, however, possible that they have a poorer verbal working memory compared to the controls (cf. Andersson & Lyxell, 2007; Siegel & Ryan, 1989; Swanson & Sachse-Lee, 2001). A verbal working-memory deficit could potentially account for some of the problems that the two MD groups displayed with mathematic word problem solving.

Future studies should take into account these limitations. In addition, an important objective for future research on children with MD is to examine multiple areas of mathematics from a longitudinal developmental perspective (cf. Jordan, Hanich, & Kaplan, 2003). This type of study is necessary in order to examine the stability of MD in children, that is, identifying children with MD whose difficulties reflect either a delay or a deficit in cognitive development (Kulak, 1993). Furthermore, by examining multiple areas of mathematics in longitudinal studies it will be possible to identify those areas that continue to be problematic for children with MD and those areas that do not. As such, it will provide a



more comprehensive test of the first hypothesis of the present study, that children with MD only and MD–RD display more severe problems with arithmetic facts retrieval, mathematic word problem solving, and written multidigit arithmetic calculation compared to areas such as place value and calculation principles.

A longitudinal research design assessing multiple areas of mathematics will also present an opportunity to further examine whether the poor performance of the two MD groups on mathematic word problem solving and multidigit written calculation is due to weaknesses in basic mathematical skills such as arithmetic fact retrieval, understanding of calculation principles, and computation (i.e., Hypotheses 3 and 4). In addition, to examine potential mathematic problems for RD-only children future studies should include more linguistically demanding word problem-solving tasks. The inclusion of more linguistically demanding tasks should also provide a test of the hypothesis that children with MD–RD have a disadvantage compared to children with MD only on tasks that are mediated by language. Further studies are also needed to expand researchers' knowledge regarding the time-telling skills of children with MD. For example, this skill should be examined in younger and older children with MD to determine the order of skill acquisition and the age at which MD children catch up with their normally achieving peers.

## Conclusions

The present study provides further empirical evidence that children with MD or MD–RD have weaknesses in multiple areas of mathematics. For example, a new and important finding is that children with MD only and MD–RD have substantial problems with telling time. Place value, in contrast, is the only domain that did not differentiate children with MD from the control children. Deficits in fact retrieval are a cardinal characteristic of children with MD. This deficit also seems to be an underlying factor to the problems children with MD display with multidigit calculation and mathematic word problem solving. The substantial difficulties that both groups with MD display with mathematic word problem solving can be attributed to several subprocesses involved in problem solving. Besides poor skills in multidigit calculation and arithmetic fact retrieval, and poor understanding of calculation principles, children with MD might have deficits related to specific problem-solving processes such as establishing a problem representation and developing a solution plan. The present study failed to demonstrate that children with MD only have an advantage over MD–RD children on mathematic word problem solving, a result that might be a consequence of the fact that the present word problem tasks imposed low linguistic demands but high calculation and problem-solving demands, and thereby eliminated the linguistic advantage of the MD-only children. The present study did not detect any signs that children with RD only have any weakness in mathematics.

## References

- Ackerman, P. T., & Dykman, R. A. (1995). Reading-disabled students with and without comorbid arithmetic disability. *Developmental Neuropsychology*, 11, 351–371.
- Andersson, U. (in press). The contribution of working memory to children's mathematical word problem solving. *Applied Cognitive Psychology*.
- Andersson, U., & Lyxell, B. (2007). Working memory deficits in children with mathematical difficulties: A general or specific deficit? *Journal of Experimental Child Psychology*, 96, 197–228.
- Ashcraft, M. H. (1992). Cognitive arithmetic: A review of data and theory. *Cognition*, 44, 75–106.
- Ashcraft, M. H. (1995). Cognitive psychology and simple arithmetic: A review and summary of new directions. *Mathematical Cognition*, 1, 3–34.
- Aunola, K., Leskinen, E., Lerkkanen, M.-K., & Nurmi, J.-E. (2004). Developmental dynamics of math performance from preschool to grade 2. *Journal of Educational Psychology*, 96, 699–713.
- Baddeley, A. D. (1996). Exploring the central executive. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 49(A), 5–28.
- Baroody, A. J., & Wilkins, J. L. (1999). The development of informal counting, number, and arithmetic skills and concepts. In J. V. Copley (Ed.), *Mathematics in the early years* (pp. 48–65). Washington, DC: National Association for the Education of Young Children.
- Bock, K., Irwin, D. E., Davidson, D. J., & Levelt, W. J. M. (2003). Minding the clock. *Journal of Memory and Language*, 48, 653–685.
- Bull, R., & Johnston, R. S. (1997). Children's arithmetical difficulties: Contributions from processing speed, item identification, and short-term memory. *Journal of Experimental Child Psychology*, 65, 1–24.
- Bull, R., & Scerif, G. (2001). Executive functioning as a predictor of children's mathematics ability: Inhibition, switching, and working memory. *Developmental Neuropsychology*, 19, 273–293.
- Case, R., Sandieson, R., & Dennis, S. (1986). Two cognitive-developmental approaches to the design of remedial instruction. *Cognitive Development*, 1, 293–333.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, 44, 1–42.
- Dehaene, S., & Cohen, L. (1991). Two mental calculation systems: A case study of severe calculia with preserved approximation. *Neuropsychologia*, 29, 1045–1074.
- Dehaene, S., Spelke, E., Pinel, P., Stanescu, R., & Tsivkin, S. (1999, May 7). Sources of mathematical thinking: Behavioral and brain-imaging evidence. *Science*, 284, 970–974.
- Dowker, A. (2005). *Individual differences in arithmetic: Implications for psychology, neuroscience and education*. Hove, England: Psychology Press.
- Friedman, W. J., & Laycock, F. (1989). Children's analog and digital clock knowledge. *Child Development*, 60, 357–371.
- Fuchs, L. S., & Fuchs, D. (2002). Mathematical problem-solving profiles of students with mathematics disabilities with and without comorbid reading disabilities. *Journal of Learning Disabilities*, 35, 563–573.
- Gathercole, S. E., Alloway, T. P., Willis, C., & Adams, A. M. (2006). Working memory in children with reading disabilities. *Journal of Experimental Child Psychology*, 93, 265–281.
- Geary, D. C. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin*, 114, 345–362.
- Geary, D. C. (1994). *Children's mathematical development: Research and applications*. Washington, DC: American Psychological Association.
- Geary, D. (2004). Mathematics and learning disabilities. *Journal of Learning Disabilities*, 37, 4–15.
- Geary, D. C., Bow-Thomas, C. C., & Yao, Y. (1992). Counting knowledge and skill in cognitive addition: A comparison of normal and mathematically disabled children. *Journal of Experimental Child Psychology*, 54, 372–391.
- Geary, D. C., & Brown, S. C. (1991). Cognitive addition: Strategy choice and speed-of-processing differences in gifted, normal, and mathematically disabled children. *Developmental Psychology*, 27, 398–406.
- Geary, D. C., Brown, S. C., & Samaranayake, V. A. (1991). Cognitive addition: A short longitudinal study of strategy choice and speed-of-

- processing differences in normal and mathematically disabled children. *Developmental Psychology*, 27, 787–797.
- Geary, D. C., Hamson, C. O., & Hoard, M. K. (2000). Numerical and arithmetical cognition: A longitudinal study of process and concept deficits in children with learning disability. *Journal of Experimental Child Psychology*, 77, 236–263.
- Geary, D. C., Hoard, M. K., & Hamson, C. O. (1999). Numerical and arithmetical cognition: Patterns of functions and deficits in children at risk for a mathematical disability. *Journal of Experimental Child Psychology*, 74, 213–239.
- Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities*, 38, 293–304.
- Ginsburg, H. P. (1997). Mathematics learning disabilities: A view from developmental psychology. *Journal of Learning Disabilities*, 30, 20–33.
- Gonzalez, J. E. J., & Espinel, A. I. G. (2002). Strategy choice in solving arithmetic word problems: Are there differences between students with learning disabilities, G-V poor performance and typical achievement students? *Learning Disability Quarterly*, 25, 113–122.
- Hanich, L. B., Jordan, N. C., Kaplan, D., & Dick, J. (2001). Performance across different areas of mathematical cognition in children with learning disabilities. *Journal of Educational Psychology*, 93, 615–626.
- Hitch, G. J., & McAuley, E. (1991). Working memory in children with specific arithmetical learning difficulties. *British Journal of Psychology*, 82, 375–386.
- Järpsten, B., & Taube, K. (1997). *Verbal ability test for grade 4–6*. Stockholm: Psychology Publishing House.
- Jordan, N. C., & Hanich, L. B. (2000). Mathematical thinking in second grade children with different forms of LD. *Journal of Learning Disabilities*, 33, 567–578.
- Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). A longitudinal study of mathematical competencies in children with specific mathematics difficulties versus children with comorbid mathematics and reading difficulties. *Child Development*, 74, 834–850.
- Jordan, N. C., Hanich, L. B., & Uberti, H. Z. (2003). Mathematical thinking and learning difficulties. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills: Constructing adaptive expertise* (pp. 359–383). Mahwah, NJ: Erlbaum.
- Jordan, N. C., Kaplan, D., & Hanich, L. B. (2002). Achievement growth in children with learning difficulties in mathematics: Findings of a two-year longitudinal study. *Journal of Educational Psychology*, 94, 586–597.
- Jordan, N. C., & Montani, T. O. (1997). Cognitive arithmetic and problem solving: A comparison of children with specific and general mathematics difficulties. *Journal of Learning Disabilities*, 30, 624–634, 648.
- Kalyuga, S., Chandler, P., & Sweller, J. (2004). When redundant on-screen text in multimedia technical instruction can interfere with learning. *Human Factors*, 46, 567–581.
- Kintsch, W., & Greeno, J. G. (1985). Understanding solving arithmetic problems. *Psychological Review*, 92, 109–129.
- Kulak, A. G. (1993). Parallels between math and reading disability: Common issues and approaches. *Journal of Learning Disabilities*, 26, 666–673.
- Lee, T. M. C., Cheung, C. C. Y., Chan, J. K. P., & Chan, C. C. H. (2000). Trail-making across languages. *Journal of Clinical and Experimental Neuropsychology*, 22, 772–778.
- Lewis, C., Hitch, G. J., & Walker, P. (1994). The prevalence of specific arithmetic difficulties and specific reading difficulties in 9- to 10-year old boys and girls. *Journal of Child Psychology & Psychiatry & Allied Disciplines*, 35, 283–292.
- Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). Oxford, England: Oxford University Press.
- Logie, R., & Baddeley, A. D. (1987). Cognitive processes in counting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13, 310–326.
- Malmquist, E. (1969). *Reading difficulties in primary school: Experimental studies*. Stockholm: Education Publishing House.
- Malmquist, E. (1977). *Reading and writing difficulties in children. Analysis and treatment*. Lund, Sweden: Gleerups.
- Mayer, R. E., & Hegarty, M. (1996). The process of understanding mathematical problems. In R. J. Sternberg & T. Ben-Zeev (Eds.), *The nature of mathematical thinking* (pp. 29–53). Mahwah, NJ: Erlbaum.
- Mazzocco, M. M. M., & Thompson, R. E. (2005). Kindergarten predictors of math learning disability. *Learning Disabilities Research & Practice*, 20, 142–155.
- McCloskey, M. (1992). Cognitive mechanisms in numerical processing: Evidence from acquired dyscalculia. *Cognition*, 44, 107–157.
- McCloskey, M., Caramazza, A., & Basili, A. (1985). Cognitive mechanisms in number processing and calculation: Evidence from dyscalculia. *Brain and Cognition*, 4, 171–196.
- McLean, J. F., & Hitch, G. J. (1999). Working memory impairments in children with specific arithmetic learning difficulties. *Journal of Experimental Child Psychology*, 74, 240–260.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, 41, 49–100.
- Ostad, S. A. (1997). Developmental differences in addition strategies: A comparison of mathematically disabled and mathematically normal children. *British Journal of Educational Psychology*, 67, 345–357.
- Ostad, S. A. (1998). Developmental differences in solving simple arithmetic word problems and simple number-fact problems: A comparison of mathematically normal and mathematically disabled children. *Mathematical Cognition*, 4, 1–19.
- Parmar, R. S., Cawley, J. R., & Frazita, R. R. (1996). Word problem-solving by students with and without math disabilities. *Exceptional Children*, 62, 415–429.
- Passolunghi, M. C., & Pazzaglia, F. (2004). Individual differences in memory updating in relation to arithmetic problem solving. *Learning and Individual Differences*, 14, 219–230.
- Passolunghi, M. C., & Siegel, L. S. (2001). Short-term memory, working memory, and inhibitory control in children with difficulties in arithmetic problem solving. *Journal of Experimental Child Psychology*, 80, 44–57.
- Passolunghi, M. C., & Siegel, L. S. (2004). Working memory and access to numerical information in children with disability in mathematics. *Journal of Experimental Child Psychology*, 88, 348–367.
- Pedhazur, E. J. (1982). *Multiple regression in behavioral research*. New York: Holt.
- Raven, J. C. (1976). *Standard Progressive Matrices*. Oxford, England: Oxford Psychologists Press.
- Raven, J. C., Court, J. H., & Raven, J. (1996). *Standard Progressive Matrices (Raven manual: Section 3)*. Oxford, England: Oxford Psychologists Press.
- Russell, R. L., & Ginsburg, H. P. (1984). Cognitive analysis of children’s mathematical difficulties. *Cognition and Instruction*, 1, 217–244.
- Siegel, L. S., & Ryan, E. B. (1989). The development of working memory in normally achieving and subtypes of learning disabled children. *Child Development*, 60, 973–980.
- Siegler, R. S. (1988). Individual differences in strategy choices: Good students, not-so-good students, and perfectionists. *Child Development*, 59, 833–852.
- Siegler, R. S., & McGilly, K. (1989). Strategy choices in children’s time-telling. In I. Levin & D. Zakay (Eds.), *Time and human cognition: A life-span perspective* (pp. 185–218). Oxford, England: North-Holland.
- Swanson, H. L. (1992). Generality and modifiability of working memory among skilled and less skilled readers. *Journal of Educational Psychology*, 84, 473–488.



- Swanson, H. L. (1994). Short-term memory and working memory: Do both contribute to our understanding of academic achievement in children and adults with learning disabilities? *Journal of Learning Disabilities*, 27, 34–50.
- Swanson, H. L. (2004). Working memory and phonological processing as predictors of children's mathematical problem solving at different ages. *Memory & Cognition*, 32, 648–661.
- Swanson, H. L. (2006). Cross-sectional and incremental changes in working memory and mathematical problem solving. *Journal of Educational Psychology*, 98, 265–281.
- Swanson, H. L., & Beebe-Frankenberger, M. (2004). The relationship between working memory and mathematical problem solving in children at risk and not at risk for serious math difficulties. *Journal of Educational Psychology*, 96, 471–491.
- Swanson, H. L., & Sachse-Lee, C. (2001). Mathematical problem solving and working memory in children with learning disabilities: Both executive and phonological processing are important. *Journal of Experimental Child Psychology*, 79, 294–321.
- Sweller, J. (2005). The redundancy principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 159–167). New York: Cambridge University Press.
- Temple, C. M. (1991). Procedural dyscalculia and number fact dyscalculia: Double dissociation in developmental dyscalculia. *Cognitive Neuropsychology*, 8, 155–176.
- Vakali, M. (1991). Clock time in seven to ten year-old children. *European Journal of Psychology of Education*, 6, 325–336.

Received December 21, 2006

Revision received August 7, 2007

Accepted August 20, 2007 ■

### Members of Underrepresented Groups: Reviewers for Journal Manuscripts Wanted

If you are interested in reviewing manuscripts for APA journals, the APA Publications and Communications Board would like to invite your participation. Manuscript reviewers are vital to the publications process. As a reviewer, you would gain valuable experience in publishing. The P&C Board is particularly interested in encouraging members of underrepresented groups to participate more in this process.

If you are interested in reviewing manuscripts, please write to the address below. Please note the following important points:

- To be selected as a reviewer, you must have published articles in peer-reviewed journals. The experience of publishing provides a reviewer with the basis for preparing a thorough, objective review.
- To be selected, it is critical to be a regular reader of the five to six empirical journals that are most central to the area or journal for which you would like to review. Current knowledge of recently published research provides a reviewer with the knowledge base to evaluate a new submission within the context of existing research.
- To select the appropriate reviewers for each manuscript, the editor needs detailed information. Please include with your letter your vita. In the letter, please identify which APA journal(s) you are interested in, and describe your area of expertise. Be as specific as possible. For example, "social psychology" is not sufficient—you would need to specify "social cognition" or "attitude change" as well.
- Reviewing a manuscript takes time (1–4 hours per manuscript reviewed). If you are selected to review a manuscript, be prepared to invest the necessary time to evaluate the manuscript thoroughly.

Write to Journals Office, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

# Prediction of Children's Academic Competence From Their Effortful Control, Relationships, and Classroom Participation

Carlos Valiente, Kathryn Lemery-Chalfant, Jodi Swanson, and Mark Reiser  
Arizona State University

The authors examined the relations among children's effortful control, school relationships, classroom participation, and academic competence with a sample of 7- to 12-year-old children ( $N = 264$ ). Parents and children reported on children's effortful control, and teachers and children reported on children's school relationships and classroom participation. Children's grade point averages (GPAs) and absences were obtained from school-issued report cards. Significant positive correlations existed between effortful control, school relationships, classroom participation, and academic competence. Consistent with expectations, the teacher-child relationship, social competence, and classroom participation partially mediated the relation between effortful control and change in GPA from the beginning to the end of the school year. The teacher-child relationship and classroom participation also partially mediated the relation between effortful control and change in school absences across the year.

**Keywords:** effortful control, peer and teacher relationships, classroom participation, academic competence

Children's academic competence is central to their future success. The importance of successfully navigating the challenges of the school environment is highlighted by findings that academic competence is a significant correlate of positive mental health and high school graduation (Caspi, Elder, & Bem, 1987; Ensminger & Slusarcick, 1992). Despite the importance of school success, 15% of adults report that they have not completed high school (Stoops, 2004). Although the majority of research on school success has focused on curricula, structure, teacher-child ratios, and intelligence, there is an increased awareness of the important roles children's regulatory abilities, school-related relationships, and classroom participation play in contributing to their academic competence. Indeed, Blair (2002) noted that some longitudinal evidence indicates that social and emotional factors relate to aspects of school success or failure even when controlling for general intelligence at school entry.

The literature considering indices of children's regulatory abilities and measures of academic competence is growing, as is the literature on relational and motivational correlates of school success. However, studies that bridge these literatures are rare. The current study begins to fill this gap and was designed (a) to test if

effortful control (an index of regulatory abilities) predicts changes in academic competence (i.e., grades and absences) across a school year; (b) to test if students' relationships with teachers and peers, as well as their classroom participation, predict changes in academic competence; and (c) to test if relationships and classroom participation partially mediate the relation between effortful control and academic competence. Simultaneously considering constructs from traditionally different areas of research may clarify if and how children's regulatory abilities predict their academic competence.

We used effortful control (EC) as an index of children's regulatory abilities. EC is defined as "the efficiency of executive attention—including the ability to inhibit a dominant response and/or to activate a subdominant response, to plan, and to detect errors" (Rothbart & Bates, 2006, p. 129). Children high in EC are believed to be able to voluntarily control their attention and behavior as needed. EC is measured in a variety of ways, often with measures of attentional regulation, persistence, and the ability to delay gratification, as well as with indices of the ability to voluntarily inhibit or activate behavior (Kochanska, Murray, & Harlan, 2000; Rothbart & Bates, 2006). EC processes are linked to children's emotion-related regulation, and they modulate emotional reactivity and behaviors (Rothbart & Bates, 2006).

---

Carlos Valiente, Jodi Swanson, and Mark Reiser, School of Social and Family Dynamics, Arizona State University; Kathryn Lemery-Chalfant, Department of Psychology, Arizona State University.

This research was funded in part by a grant from the Faculty Grant in Aid program, Arizona State University, and from National Science Foundation CAREER Award BCS-0546096 to Carlos Valiente. We thank the principals, teachers, and students of the Casa Grande School District for their support of this research. We also thank Adriana Umaña-Taylor and Sandra Simpkins for their comments on a draft of this article.

Correspondence concerning this article should be addressed to Carlos Valiente, School of Social and Family Dynamics, Arizona State University, Tempe, AZ 85287-3701. E-mail: valiente@asu.edu

## Relations Between Children's Effortful Control and Academic Competence

Several investigators have argued that emotional competence and processes involving executive attention are important for academic success (Blair, 2002; Raver, 2002). Huffman, Mehlinger, and Kerivan (2000) hypothesized that children's regulatory abilities contribute to competence beyond measures of IQ. In one study, 60% of teachers reported that being sensitive and not disruptive represented important aspects of academic readiness (Lewit & Baker, 1995). Children high in EC likely have many of



these skills, do not easily divert from tasks (Zimmerman, 1998), and process detailed situations more accurately than do their peers low in EC (Lemerise & Arsenio, 2000; NICHD Early Child Care Research Network, 2003).

Some evidence supports the hypothesis that components of EC (e.g., attentional regulation, persistence, or delay of gratification) are positively related to reading, math, and linguistic abilities as well as teachers' reports of competence (Fabes, Martin, Hanish, Anders, & Madden-Derdich, 2003; NICHD Early Child Care Research Network, 2003), although EC sometimes does not relate to grade retention (Willson & Hughes, 2006). Findings that preschoolers' delay of gratification predicts future verbal intellectual ability and SAT scores provide some evidence that the relations between regulatory abilities and academic competence persist over time (Rodriguez, Mischel, & Shoda, 1989; Shoda, Mischel, & Peake, 1990). Moreover, some data indicate that the relations between EC and academic competence are similar cross-ethnically. For example, Valiente, Lemery-Chalfant, and Castro (2007) found that Mexican American children's EC was related to teacher-reported academic competence and absenteeism.

The preceding review indicates that children's EC is positively related to measures of academic competence; however, not all findings are consistent or significant effects are reduced in magnitude when control variables (e.g., sex or socioeconomic status; SES) or correlates of academic competence are added to the models. In addition, the relations generally account for a modest amount of the variance, suggesting that other constructs are also important when considering academic competence. On the basis of the broader literature, it seems likely that students' relationships and classroom participation may mediate the relations between EC and academic competence.

### Relations Between Children's Relationships and Academic Competence

As noted above, in addition to EC, the relationships children develop and maintain in school have been associated with their academic competence. A supportive teacher-child relationship may buffer children from some risk factors associated with poor performance, perhaps because teachers are more likely to provide extra assistance to children with whom they have a positive relationship (Resnick et al., 1997). Consistent with this hypothesis, Hamre and Pianta (2001) predicted that a high-quality teacher-child relationship motivates teachers to invest extra resources that can promote children's school success. In contrast, a conflictual teacher-child relationship may increase stress for the child that may interfere with learning and motivation. Longitudinal data suggest that declines in the nurturant teacher-child relationship precede declines in achievement, and there is evidence that teacher-reported negativity in the teacher-child relationship is related to achievement test scores even when controlling for verbal IQ (Hamre & Pianta, 2001; Midgley, Feldlaufer, & Eccles, 1989).

Positive relationships with peers also contribute to children's academic achievement (Raver, 2002). Indeed, components of social competence such as peer acceptance and friendships are hypothesized to promote social inclusion in the classroom, which may yield resources that foster interpersonal and academic success (Ladd, 2003). Welsh, Parke, Widaman, and O'Neil (2001) found that positive social skills were associated with academic compe-

tence, and evidence supports the premise that peer acceptance and general levels of social competence are linked to performance in math, reading, and language (see Ladd, 2003, for a review).

The preceding review suggests that both children's EC and relationships are related to academic competence. In addition, some data support the hypothesis that EC is related to children's relationships, and this is necessary for relationships to mediate the relation between EC and academic competence. When children are low in EC and disruptive, they are at increased risk of developing poor relationships with their teachers and receiving low levels of instruction (Berndt & Keefe, 1995; Birch & Ladd, 1997; Murray & Greenberg, 2000). This may contribute to negative perceptions of the classroom, which interfere with motivation for subsequent learning and performance (Wentzel, 1999). Consistent with this line of reasoning, children who lack social skills are viewed as difficult to teach and receive low levels of positive feedback (Arnold, Homrok, Ortiz, & Stowe, 1999; McEvoy & Welker, 2000; Shores & Wehby, 1999), but studies that directly measure EC and examine the hypothesized relations are needed.

There is support for the hypothesis that EC is positively related to indices of social competence. For example, children high in EC are often rated high in compliance (Kochanska, Coy, & Murray, 2001; Kochanska, Murray, & Coy, 1997), sympathy (Eisenberg & Fabes, 1998; Valiente et al., 2004), and social competence (Eisenberg, Gershoff, et al., 2001; see Rothbart & Bates, 2006, for a review). In contrast, children low in EC likely have difficulty modulating negative emotions, complying with others' requests, and avoiding conflictual peer interactions. In summary, there is evidence that academic competence is associated with both EC and children's relationships, but it is not clear if EC provides unique prediction of academic competence or if relationships partially mediate the EC and academic competence associations.

### Relations Between Children's Classroom Participation and Academic Competence

Some findings indicate that students' classroom participation is associated with their grades and absences and that classroom participation might partially mediate the relations between EC and academic competence. Theorists argue that participation may reflect an internal motivation and learning-goal orientation that directs one's behavior toward classroom tasks and demands (Dweck, 1989; Finn, 1993; Gottfried, Fleming, & Gottfried, 1994). In a national report on educational statistics, Finn (1993) noted that students who rarely participate in their classrooms are at risk to perform poorly beyond risks associated with race, ethnicity, language, or family income. Furthermore, scholars suggest that motivation contributes to academic outcomes because it directs students' actions and activities, perhaps because they are motivated to pursue goals valued in the school context (Wentzel, 1999; Wigfield, Eccles, Schiefele, Roeser, & Davis-Kean, 2006). Evidence indicates that measures of engagement such as classroom participation and school liking are related to academic progress, math skills, language skills, and attendance (Ladd, Birch, & Buhs, 1999; Valiente et al., 2007). Children low in engagement are likely to have difficulty following rules and capitalizing on learning opportunities that are often correlated with indices of cognitive functioning (Bronson, Tivnan, & Seppanen, 1995; Hughes & Kwok, 2006).



After reviewing the motivation-to-succeed literature, Eccles, Wigfield, and Schiefele (1998) noted that "the highest priority in this area is attention to the influence of emotions on motivation" (p. 1075). Although progress has been made (Wigfield et al., 2006), Eisenberg (2006) suggested in the introduction to the *Handbook of Child Psychology* that much more work is necessary. Relations between EC and classroom participation might be expected, because children must modulate emotions and demands that occur in school environments to remain engaged (Alexander & Entwisle, 1988). Consistent with this line of reasoning, Valiente et al. (2007) found that EC was positively related to school liking (an index of engagement), and school liking mediated the relation between EC and academic competence.

### The Present Study

Findings from several often disparate literatures support the premise that children's EC, relationships with teachers and peers, and classroom participation represent important correlates of academic competence. However, it is unclear if these variables offer unique or overlapping prediction of academic competence. Our first aim in this study was to test if EC predicted measures of academic competence (i.e., grades and absences) at the conclusion of a school year (i.e., spring) while controlling for academic competence at the beginning of the school year (i.e., fall). Our second goal was to test if prediction from EC remained significant after adding measures of children's relationships and a measure of classroom participation to the model. We expected positive zero-order relations among children's EC, the teacher-child relationship, social competence, classroom participation, and GPA; the reverse pattern of relations was expected for school absences. In an effort to extend the literature and on the basis of some empirical evidence (Valiente et al., 2007) and theory (Eisenberg, Sadovsky, & Spinrad, 2005), we predicted that the teacher-child relationship, social competence, and classroom participation would mediate the relations between EC and academic competence.

The extant literature indicates that measures of children's SES often relate to children's performance in school and attendance. For example, income was positively related to both the classroom environment (NICHD Early Child Care Research Network, 2005) and students' achievement (Davis-Kean, 2005). In a review, Haveman and Wolfe (1995) concluded that poverty limits students' academic competence and academic years completed. Nevertheless, as noted above, low classroom participation places a child at risk for school failure regardless of family income (Finn, 1993). To examine the influence of SES, we added SES to all models and expected SES to be positively related to grades and negatively related to absences.

The literature also suggests that children's sex is associated with their academic competence; however, findings regarding sex are somewhat less consistent than findings for SES. Whereas sex differences in math and reading sometimes fail to reach significance (Davis-Kean, 2005; Simpkins, Davis-Kean, & Eccles, 2006), some research suggested that boys outperformed girls on math tasks (Frome & Eccles, 1998; Jordan, Kaplan, Oláh, & Locuniak, 2006), and other reports found that girls achieved better reading performance than boys (Frome & Eccles, 1998). Although sex differences such as these are generally small, because of these

relations, we added children's sex to all equations when testing our hypotheses.

In addition, because 47% of the sample in this study was Mexican American and the next largest ethnic group represented was European American (30%), we tested if the strength of the relations differed for Mexican American and European American children. Findings from previous work indicate that Latino students perform more poorly in reading and math than do their European American peers (Children's Action Alliance, 1999; U.S. Department of Education, 2004).<sup>1</sup> Latino students also tend to be absent, to be tardy, and to drop out of school prior to graduation more often than students of other ethnicities (Finn, 1993; U.S. Bureau of the Census, 2004).

Despite some mean level differences, and although much of the data in this area stems from European American children, it seems likely that EC, school relationships, and classroom participation are also important correlates of Mexican American children's academic competence. There is some evidence that prediction of developmental outcomes from measures of EC in other cultures (e.g., Indonesia and China) is similar to prediction of outcomes found in U.S. samples (Eisenberg, Pidada, & Liew, 2001; Zhou, Eisenberg, Wang, & Reiser, 2004). Furthermore, the relations between U.S. minority children's regulatory abilities and their socioemotional and school functioning are similar to relations found in studies that include mostly European American children (Blair, 2002; Gilliom, Shaw, Beck, Schonberg, & Lukon, 2002; Schultz, Izard, & Ackerman, 2000). Thus, we did not expect the relations among the study variables to be different for Mexican American versus European American children.

### Method

#### *Procedure and Participants*

Participants were recruited from two schools in a southwestern U.S. city. Before the study began, parents ( $N = 561$ ) received an introductory letter informing them of enrollment and participation procedures. In addition, research assistants were available in the schools during parent-teacher conferences to enroll parents and answer questions about the study. To increase the reliability of the construct scores and to reduce shared method variance, we used a multireporter (child, primary caregiver parent, and teacher) method to assess the key constructs. The questionnaires sent to parents assessed children's EC and social competence. Parent packets were available in either English or Spanish, because some parents ( $n = 40$ ) preferred to complete their packets in Spanish. A translation and back-translation method was used, and the original English version was compared with the back-translated version to determine equivalence. The process of translation and back translation continued until all differences were resolved. Teachers ( $N = 22$ ; teachers reported on an average of 12 children) reported on the teacher-child relationship, children's social competence, and classroom participation. Children reported on their EC, the teacher-child relationship, and their classroom participation. Dur-

<sup>1</sup> We use *Latino* when describing the pan-ethnic population that comprises those of Spanish ethnicity and *Mexican American* when either describing our sample (composed only of Mexican American Latinos) or when referring only to the Mexican American subethnic Latino group.



ing the school day, a research assistant read all items to children in their classroom. All questionnaires were completed between March and April, and participants were compensated for their participation.

Parents of 122 boys and 142 girls (47% of those eligible) provided consent for themselves and for their children to participate. Children were between the ages of 7 and 12 years ( $M = 9.57$  years,  $SD = 1.04$ ), attended 1 of 22 regular education classrooms, and provided assent. Forty-seven percent of the parent and child participants were Mexican American, 30% were European American, 5% were African American, 8% were Native American, and 10% were of other ethnic origins. To reduce the heterogeneity of the sample, we did not include children who were solely enrolled in special education services and did not spend any time in regular education classrooms in this study.

The sample represents the sex and ethnic population composition of the classrooms (i.e., the population of the classrooms was 48% boys and 52% Mexican American, 34% European American, 8% African American, and 6% Native American children), demonstrating that our sample of children closely resembles all of those eligible. To further compare those children whose parents provided consent with those whose parents did not, we examined the absence rate, a percentage rate published by the Arizona Department of Education, of those who participated with those eligible. The absence rate in our sample (4%) closely represents the absence rate for those eligible (5%). Because public access of GPA data is unavailable, we were not able to conduct a similar comparison for GPA.

Children were predominantly from two-parent homes (80%) in which the primary caregiver was the child's biological mother (84%). Seventeen percent of primary caregivers (percentages for secondary caregivers [75% were fathers] are in parentheses; 17%) had less than a high school education, 26% (37%) had a high school education, 28% (19%) had some college education, 16% (11%) had a 2-year-college or trade degree, 7% (7%) had a 4-year-college degree, and 6% (9%) had attended graduate school. Family income ranged from below \$15,000 to above \$150,000 per year and had a mean range of \$30,000 to \$50,000 per year.

## Measures

**SES.** Income, primary caregivers' education, and secondary caregivers' education were highly related ( $rs > .45$ ,  $p < .01$ ); therefore, after standardizing, the items were averaged to form a measure of SES ( $\alpha = .76$ ).

**Effortful control.** Children and parents reported on children's EC. Items from the Attention Shifting (e.g., "I am [Your child is] good at keeping track of several different things that are happening around me [him/her]"), Activation Control (e.g., "If I have [Your child has] a hard assignment to do, I get [he/she gets] started right away"), and Inhibitory Control (e.g., "When someone tells me [your child] to stop doing something, it is easy for me [him/her] to stop") subscales from the Early Adolescent Temperament Questionnaire (Capaldi & Rothbart, 1992) assessed children's EC. Items were rated on a 5-point Likert-type scale (1 = *almost always untrue* to 5 = *almost always true*), and the average of the items (18 for parents and 16 for children) served as a composite of parent-reported and child-reported EC ( $\alpha s = .89$  and  $.73$  for parents and children, respectively). Capaldi and Rothbart (1992) reported that

scores from the scale have good discriminant validity, and there is evidence that parent-reported EC and observed indices of EC are significantly related (Eisenberg, Gershoff, et al., 2001; Valiente et al., 2003).

**Teacher-child relationship.** Teachers reliably ( $\alpha = .90$ ) rated 20 items on a 5-point scale (1 = *definitely not true* to 5 = *definitely true*) from the Student-Teacher Relationship Scale to report on the closeness and conflict of the teacher-child relationship (e.g., Hamre & Pianta, 2001; Pianta, 2001). This scale has been used often in studies with elementary school children, and scores have demonstrated internal consistency. Children rated their relationship with their teacher on a 3-point scale ( $\alpha = .92$ ; 1 = *not at all* to 3 = *a lot of the time*) using an age-appropriate version of the Student-Teacher Relationship Scale. Evidence that the measure correlates in the expected directions with later academic performance, attitudes, and involvement supports the convergent validity of the Student-Teacher Relationship Scale (Birch & Ladd, 1997; Hamre & Pianta, 2001).

**Social competence.** Two subscales of a slightly modified (Eisenberg, Fabes, Guthrie, & Reiser, 2000) version of Harter's (1982) Perceived Competence Scale for Children were used to assess children's social competence. Socially appropriate behavior was the average of four items (e.g., "This child is usually well behaved";  $\alpha s = .87$  for parents and  $.85$  for teachers). Popularity was the average of three items (e.g., "This child has a lot of friends";  $\alpha s = .88$  for parents and  $.89$  for teachers). Parents' reports of socially appropriate behavior and popularity were highly correlated,  $r(217) = .73$ ,  $p < .001$ , as were teachers' reports,  $r(236) = .53$ ,  $p < .001$ . Thus, the scales were averaged within reporter to form separate measures of parents' and teachers' reports of social competence.

**Classroom participation.** Teachers used 11 items from the Teacher Rating Scale of School Adjustment (Birch & Ladd, 1997; Ladd, Kochenderfer, & Coleman, 1996) to rate children's classroom participation. Items were rated on a 3-point scale (0 = *doesn't apply* to 2 = *certainly applies*). Teachers' reports of children's classroom participation (e.g., "This child follows instructions," "This child challenges him/herself to do well in school") were reliable ( $\alpha = .94$ ). Children reliably ( $\alpha = .67$ ) rated items on an age-appropriate version (e.g., "I follow my teacher's instructions") of this measure.

**Academic competence.** Official school records were used to obtain measures of children's academic competence. At the conclusion of the school year, we obtained records of full school days missed and tardies from the fall and spring quarters of the school year. Because the number of full school days missed and tardies were significantly related at the first and last quarters ( $rs = .24$  and  $.27$ ,  $ps < .001$ , respectively), we standardized the number of full school days missed and tardies and then averaged the standardized scores. In the remainder of the article, we refer to this composite as *absences*. Consistent with Pierce, Hamm, and Vandell (1999), we averaged scores in language, vocabulary, and math (all  $rs > .60$ ) to form children's fall and spring GPAs (1 = *a grade of F* to 5 = *a grade of A*).

## Results

Prior to hypothesis testing, we computed a series of preliminary analyses to test for potential age and sex differences. Next, we

examined zero-order relations among the study variables. Finally, mixed model regressions were computed to test the hypotheses. We concluded by testing if the strength of the findings was different for Mexican American versus European American participants.

Complete data were available for 77% of the participants. To avoid problems associated with listwise deletion (see Schafer & Graham, 2002), we imputed missing values using the expectation maximization algorithm after specifying a normal distribution with the missing value analysis program in SPSS Version 12.0. Little's missing completely at random test was not significant,  $\chi^2(288) = 291.60$ , *ns*, which supports this method of imputing missing data. Because the patterns of findings were similar for imputed and nonimputed data, we present the results obtained on the single imputed set.

### Preliminary Analyses

Table 1 contains the means and standard deviations for the study variables. To examine sex differences, we computed separate multivariate analyses of variance (MANOVAs) by reporter. There were significant multivariate effects (Hotelling's *T*) for child-reported measures,  $F(3, 260) = 5.86$ ,  $p < .001$ , and for teacher-reported measures,  $F(2, 261) = 18.17$ ,  $p < .001$ . Univariate tests indicated that teachers reported closer teacher-child relationships with girls than with boys and rated girls higher than boys in classroom participation,  $F_s(1, 262) = 25.02$  and  $34.22$ ,  $p_s < .001$ , respectively. Girls reported higher levels of EC, closer teacher-child relationships, and more classroom participation than boys reported,  $F_s(1, 262) = 13.56$ ,  $9.36$ , and  $5.05$ ,  $p_s < .001$ ,  $.01$ , and  $.05$ , respectively. In addition, parents rated girls higher in EC than they did boys,  $t(263) = 21.60$ ,  $p < .001$ . There was also a significant multivariate effect for grades and absences,  $F(4, 259) = 7.26$ ,  $p < .001$ . Univariate tests indicated that girls per-

formed significantly better academically than boys did in fall and in spring,  $F_s(1, 262) = 21.82$  and  $24.94$ ,  $p_s < .001$ , respectively; however, there were no significant sex differences in fall or spring absence patterns.

Table 2 presents the correlations among all the variables and illustrates the similar pattern of findings for imputed versus non-imputed data. As shown in Table 2, when considering the imputed data, 63 of 78 correlations were significant. Irrespective of reporter, the measures of EC, the teacher-child relationship, social competence, and classroom participation were all significantly related to children's GPAs in fall and spring. EC, both child and parent reported, was positively related to children's social competence (parent and teacher reported), teacher-child relationship (child and teacher reported), and classroom participation (child and teacher reported). The indices of children's EC, teacher-child relationship, and social competence were significantly related to children's absences from school. The zero-order correlations provide initial support for the hypotheses.<sup>2</sup>

Table 2 also contains the within-construct relations. Consistent with previous research, child and parent reports of children's EC correlated .41, and teacher and parent reports of social competence correlated .29. In addition, teacher and child reports of the teacher-child relationship were correlated .31, and their reports of classroom participation were correlated .32. Therefore, because reports of the same construct across reporters were always significant ( $p_s < .01$ ) and because significant relations were found across reporters, to reduce the number of analyses, we created composite scores by averaging across reporters.<sup>3</sup> The remainder of the analyses use these composites.

### Regression Analyses

The observations in the present study are clustered (i.e., children are nested within classrooms); thus, prior to testing our hypotheses, we examined the intraclass correlations associated with each model. Clustered data can result in a correlation among responses from the same classroom, and standard errors and consequent significance tests may be biased if the correlation is ignored. Hox (2002) considered intraclass correlations values of .05, .10, and .15 to be small, medium, and large, respectively, but also demonstrated that even intraclass correlations of .10 can bias results. The intraclass correlations in this study ranged from .02 to .14, and the average intraclass correlation was .08 (see Tables 3 and 4 for the intraclass correlation associated with each mixed model regression). Although the intraclass correlations were generally small, because clustering can potentially bias significance tests and resulting conclusions, the remainder of the analyses were computed using mixed models in SPSS Version 12.0, with classroom as a random effect.

The variables did not exceed West, Finch, and Curran's (1995) cutoffs for skewness, kurtosis, and outliers. According to Cook's (1977) distance, there were no multivariate outliers. In each anal-

Table 1  
Descriptive Statistics for Study Variables

Variable	<i>M</i>	<i>SD</i>	Range
Socioeconomic status	3.28	1.37	0-7
Effortful control: Child report	3.41	0.52	2.00-5.00
Effortful control: Parent report	3.27	0.60	1.11-4.72
Social competence: Teacher report	3.05	0.68	1.00-4.00
Social competence: Parent report	3.00	0.78	1.00-4.06
Teacher-child relationship: Child report	2.21	0.44	1.00-3.00
Teacher-child relationship: Teacher report	3.90	0.61	1.50-5.00
Classroom participation: Child report	2.34	0.31	1.20-3.00
Classroom participation: Teacher report	2.44	0.51	1.18-3.00
GPA (fall)	3.92	0.87	1.00-5.00
GPA (spring)	3.89	0.94	1.00-5.00
Absences (fall)	0.97	1.12	0.00-6.50
Absences (spring)	1.86	1.78	0.00-12.00

Note. Statistics for socioeconomic status and absences are presented prior to standardizing scores. The standardized scores are used in all other analyses.

<sup>2</sup> Neither parent nor child reports of their own social desirability significantly correlated with the other measures.

<sup>3</sup> Because teacher and child reports of the teacher-child relationship were on a different scale, we standardized the scores before averaging the two scores. All other measures were on the same scale, so we did not standardize scores prior to creating the composites.



Table 2  
Zero-Order Correlations for Study Variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Socioeconomic status	—	.22**	.14*	.14*	.35**	-.02	.13	.11	.17*	.27**	.28**	-.07	.05
2. Effortful control: Child report	.26**	—	.39**	.33*	.26**	.32**	.34**	.32**	.42**	.36**	.41**	-.13*	-.17*
3. Effortful control: Parent report	.16*	.41**	—	.35**	.26**	.27**	.20**	.16*	.43**	.46**	.49**	-.08	-.09
4. Social competence: Teacher report	.18*	.35**	.37**	—	.28**	.19**	.65**	.15*	.77**	.38**	.46**	-.16*	-.24**
5. Social competence: Parent report	.36**	.27**	.27**	.29**	—	.05	.24**	.16*	.30**	.32**	.32**	.06	.13*
6. Teacher–child relationship: Child report	-.03	.33**	.29**	.20**	.05	—	.30**	.51**	.27**	.22**	.30**	-.11	-.17*
7. Teacher–child relationship: Teacher report	.14*	.35**	.19**	.65**	.25**	.31**	—	.26**	.70**	.29**	.33**	-.13*	-.21**
8. Classroom participation: Child report	.12*	.34**	.16*	.16*	.17**	.52**	.28**	—	.30**	.22**	.25**	-.07	-.15*
9. Classroom participation: Teacher report	.20*	.43**	.45**	.78**	.30**	.29**	.70**	.32**	—	.56**	.61**	-.16*	-.28*
10. GPA (fall)	.29*	.38**	.48**	.43**	.32**	.27**	.33**	.26**	.59**	—	.65**	-.12*	-.08
11. GPA (spring)	.27*	.42**	.52**	.48**	.31**	.30**	.35**	.27**	.63**	.67**	—	-.15*	-.23**
12. Absences (fall)	-.10	-.14*	-.08	-.15*	.09	-.10	-.13*	-.08	-.16*	-.12*	-.14*	—	.42**
13. Absences (spring)	.04	-.17**	-.11†	-.24**	.15*	-.17**	-.21**	-.16**	.28**	-.13*	-.23**	.42**	—

Note. Correlations with imputed data are below the main diagonal. Correlations based on pairwise deletion are above the main diagonal.  
†  $p < .10$ . \*  $p < .05$ . \*\*  $p < .01$ .

ysis, we controlled for children’s sex and family SES. In addition, when predicting spring GPA (or spring absences), we controlled for fall GPA (or fall absences) to examine the relation of the predictors to change in academic competence across the school year.

To address the first goal of predicting indices of academic competence from EC, we computed two mixed model regressions (see Table 3). Consistent with expectations, EC was positively related to spring GPA and negatively related to spring absences. To test the prediction that the teacher–child relationship, social competence, and classroom participation partially mediated the relation between EC and GPA (or absences), we computed additional mixed model regressions on the basis of the guidelines outlined by MacKinnon, Lockwood, Hoffman, West, and Sheets (2002). For mediation to be present, EC should predict the medi-

ator (e.g., teacher–child relationship, social competence, or classroom participation) and the mediator should predict the outcome (e.g., GPA or absences) when EC is included in the model. Full mediation exists if the relation between EC and the outcome is zero when the mediator is included in the model, whereas partial mediation exists if EC continues to predict the outcome when the mediator is in the model.

The first column of betas in Table 4 illustrates that the relation between EC and the teacher–child relationship was significant and that the teacher–child relationship was positively related to GPA and absences beyond the effects of EC (and fall GPA or absences, sex, and SES). To test for mediation and to accommodate the nonnormal distribution of the indirect effects (e.g., the product of the coefficient from the independent variable to the mediator and the coefficient from the mediator to the dependent variable are generally nonnormally distributed), we used a confidence interval method (MacKinnon et al., 2002). The upper and lower confidence limits are based on the product of the two random variables from tables produced by Meeker, Cornwell, and Aroian (1981). When the confidence intervals do not include zero, mediation is significant. As shown in Table 5, the confidence limits for mediation by the teacher–child relationship of the EC to academic competence (GPA and absences) relation do not include zero. Table 5 also indicates that the teacher–child relationship mediated 14% of the effect of EC on GPA and 40% of the effect of EC on absences.

Findings in the second panel of Table 4 demonstrate that EC was positively related to social competence and that social competence partially mediated the relation between EC and GPA but not absences (see Table 5 for the confidence limits and the percentage of variance that was mediated). In contrast, consistent with the findings for the teacher–child relationship, there was evidence that classroom participation partially mediated the relation be-

Table 3  
Prediction of GPA and Absences From EC

Predictor	$\beta$	SE $\beta$	$R^2$	ICC
Predicting spring GPA from EC				.07
Fall GPA	.54***	.06		
Sex	.14†	.08		
SES	.11*	.06	.47	
EC	.51**	.10	.51	
Predicting spring absences from EC				.02
Fall absences	.43***	.06		
Sex	.01	.09		
SES	.12*	.06	.19	
EC	-.23*	.09	.21	

Note. EC = children’s effortful control; SES = family socioeconomic status; ICC = intraclass correlation. Betas are unstandardized.  
†  $p < .10$ . \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

Table 4  
Mixed Model Regressions

Predictor	$\beta$	SE $\beta$	$R^2$	ICC	Predictor	$\beta$	SE $\beta$	$R^2$	ICC	Predictor	$\beta$	SE $\beta$	$R^2$	ICC
Predicting the teacher-child relationship from EC				.11	Predicting social competence from EC				.09	Predicting classroom participation from EC				.04
Sex	.34***	.09			Sex	.13*	.06			Sex	.13***	.04		
SES	-.01	.06	.09		SES	.18***	.04	.16		SES	.05*	.02	.17	
EC	.61***	.10	.21		EC	.48***	.07	.28		EC	.32***	.04	.32	
Predicting spring GPA from EC and the teacher-child relationship				.14	Predicting spring GPA from EC and social competence				.14	Predicting spring GPA from EC and classroom participation				.14
Fall GPA	.52***	.06			Fall GPA	.51***	.06			Fall GPA	.45***	.06		
Sex	.10	.08			Sex	.12	.08			Sex	.08	.08		
SES	.11*	.06	.47		SES	.08	.06	.47		SES	.09	.05	.47	
EC	.45***	.10	.52		EC	.44***	.10	.52		EC	.38***	.10	.51	
Teacher-child relationship	.12*	.06	.53		Social competence	.20*	.08	.53		Classroom participation	.66***	.14	.56	
Predicting spring absence from EC and the teacher-child relationship				.03	Predicting spring absences from EC and social competence				.02	Predicting spring absences from EC and classroom participation				.02
Fall absence	.41***	.06			Fall absences	.42***	.06			Fall absences	.41***	.06		
Sex	.06	.09			Sex	.01	.09			Sex	.08	.09		
SES	.11†	.06	.19		SES	.12	.06	.19		SES	.14*	.06	.19	
EC	-.15	.10	.21		EC	-.23*	.11	.21		EC	-.06	.11	.21	
Teacher-child relationship	-.15**	.06	.23		Social competence	-.01	.08	.21		Classroom participation	-.55***	.15	.25	

Note. EC = children's effortful control; SES = family socioeconomic status; ICC = intraclass correlation. Betas are unstandardized. Sex was dummy coded such that boys = 1 and girls = 2.

†  $p < .10$ . \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

tween EC and GPA and mediated the relations between EC and absences (see the third column of betas in Table 4).

Although we did not expect findings to differ by ethnicity, we computed interactions to test if the strength of the relations differed for Mexican American versus European American students, because 47% of the sample was Mexican American and the next largest percentage of ethnic population represented was European American (30%). We added the main effect of ethnicity (i.e., Mexican American vs. European American) and the interaction of ethnicity and EC when predicting the teacher-child relationship, social competence, and classroom participation. When predicting GPA or absences, we also tested if the mediators interacted with ethnicity. None of the interaction terms or main effects of ethnicity were significant. The relatively small sample sizes for the other ethnic groups (e.g., African American students) precluded the

testing of further ethnic group comparisons. Thus, consistent with expectations, there was no evidence that the strength of the findings differed across the ethnicities we tested.

## Discussion

Our primary purpose in this study was to begin to fill a gap in the literature on the regulatory and social variables related to children's school success. Despite a recent research emphasis on the importance of academic functioning for positive developmental trajectories (Ladd, 2003; Welsh et al., 2001), 15% of U.S. students drop out of formal schooling prior to graduating from high school (Stoops, 2004). Students who perform poorly often develop negative attitudes and poor scholastic habits early in their school careers; thus, a better understanding of the regulatory and social factors that are related to academic achievement in early academic grades may inform intervention programs for these students. Therefore, our first goal was to test if EC was positively related to changes in children's grades and absences, and the second goal was to test if part of the relation between EC and academic competence was mediated by children's social relationships and classroom participation.

Results supported the hypothesis that EC was positively related to grades and negatively related to absences. This relation is consistent with limited theory and data (Eisenberg et al., 2005; Hill & Craft, 2003; Raver, 2002; Valiente et al., 2007). Blair (2002) has argued that EC, and particularly attentional regulation, is related to academic competence because students who have difficulty directing their attention and behavior likely experience significant challenges when trying to learn and focus on educational material. This explanation has roots in the cognitive literature (Ruff & Rothbart, 1996) and is supported by findings that children who have diffi-

Table 5  
Confidence Limits for Mediated Effects

Effect	Lower confidence limit	Upper confidence limit	Total effect that was mediated
EC > TCR > GPA	.01	.15	14%
EC > TCR > Absence	-.18	-.02	40%
EC > SC > GPA	.02	.18	19%
EC > SC > Absence	-.08	.07	1%
EC > PART > GPA	.12	.32	36%
EC > PART > Absence	-.29	-.08	76%

Note. EC = children's effortful control; TCR = teacher-child relationship; GPA = grade point average; SC = social competence; PART = classroom participation. Bolded effects are significant at the .05 level because the confidence limits do not contain zero.



culty with attention often have poor reading and language skills (McGee, Partridge, Williams, & Silva, 1991; Tamis-LeMonda & Bornstein, 1989), perhaps because frequently shifting attention and moving between tasks interfere with both learning and completing tasks.

Consistent with our hypotheses, there was evidence that some of the relation between EC and grades (as well as absences) was mediated by the teacher-child relationship, social competence (for grades only), and classroom participation. To our knowledge, this is the first study to find that the teacher-child relationship and social competence mediate part of the relation between EC and grades. A number of investigators have found links between children's EC and their social competence, problem behaviors, and social skills (see Rothbart & Bates, 2006, for a review), and data from this study suggest that part of the reason children high in EC perform well in school is indirect and through their social relationships at school. These results support the hypothesis that students' relationships in the school context are important for school success (Baumeister & Leary, 1995; Furrer & Skinner, 2003). It is also possible that when children are low in EC and disruptive in class, they receive less classroom support from teachers and peers, miss out on learning opportunities, and view the classroom environment negatively and as something to be avoided.

In addition to the teacher-child relationship and social competence, classroom participation partially mediated the relation between EC and GPA and mediated the relation between EC and absences. These results are consistent with findings that classroom participation is positively related to math and language skills (Buhs & Ladd, 2001) and improvements in academic achievement (Ladd, Buhs, & Seid, 2000) and with evidence that school liking mediates the relation between EC and academic competence (Valiente et al., 2007). The data presented here coincide with Wentzel's (1999) hypothesis that children who are comfortable and engaged at school may also perform well academically because they are motivated to pursue goals valued in the school context.

There is increasing evidence that indices of EC are related to students' academic competence. In this article, we have argued that part of the reason for this relation is that EC provides students with both relational and motivational advantages that help them perform well. Because we found evidence of partial mediation (especially for grades), there is some evidence that EC also has a direct effect on academic competence. Perhaps there are components of EC that are more closely tied to academic performance and that are independent of relational and motivational processes. In future work, it would be useful to measure various components of EC and related constructs to more closely assess why there are both direct and indirect effects. More cognitively oriented components of EC such as planning and attention allocation may be directly related to academic competence. Inhibitory components are necessary for desirable behavior, and these may be aspects of EC that are mediated by constructs such as social competence, the teacher-child relationship, and classroom participation. One could test the working hypothesis that social and motivational processes mediate the relational and inhibitory components of EC, but the attentional advantages directly relate to academic competence by obtaining purer measures of the components of EC. Advancing this approach is one way to more fully explain why preschoolers' ability to delay gratification is associated with later verbal and quantitative SAT scores (Shoda et al., 1990).

This study demonstrates several strengths. First, we incorporated data from multiple reporters for all variables (i.e., parents, teachers, and children reported on the same variables) to reduce common source variance. Second, although researchers have recently attended to the influence of EC on academic competence, few have examined processes or mechanisms underlying this relation. A more precise understanding of why relations emerge between EC and academic performance and school absence is useful for promoting children's positive development. Third, the inclusion of a large percentage of Mexican American participants strengthens this study: The U.S. Bureau of the Census (2004) reported that by 2050, 25% of students will be of Latino descent, yet research on the normative academic functioning of this population is rare. Understanding school success among ethnic minority students is particularly important because a robust association between ethnic minority status and the likelihood of failing to complete high school has been established, with Latino students ranked most likely to drop out (Kaufman, Alt, & Chapman, 2004; U.S. Bureau of the Census, 2004). In this study, results did not differ for the European American participants and the Mexican American participants.

Finally, although not all variables were assessed longitudinally, academic competence was examined at two time points. Prediction of children's academic competence in the spring was examined while controlling for their academic competence in the fall. By controlling for fall grades or absences when examining regulatory and social contributors to children's spring grades or absences, one can assess how these factors related to academic competence beyond children's preexisting academic ability.

Despite strengths, this study had some limitations. First, we used concurrent assessments of children's EC, social competence, and classroom participation. Our data are thus correlational and do not allow for firm conclusions about directionality. Second, although the scores used in this study are valid and relate to observed indices of the relevant constructs (Birch & Ladd, 1997; Capaldi & Rothbart, 1992; Eisenberg et al., 2000), the data in the current study were assessed with questionnaires only, and future studies would benefit from using observational assessments. For example, Kochanska and colleagues have developed a battery of tasks to measure young children's EC, and methods are available to observationally code both student-teacher interactions and engagement (Kochanska et al., 2000, 2001; Ladd et al., 1999; Pianta, La Paro, Payne, Cox, & Bradley, 2002). These methods will be especially useful tools to elaborate on the beginning stages from this line of research. Finally, this line of research would be strengthened by adding a measure of IQ to the models. Evidence indicates that measures of children's regulatory abilities, relationships, and participation relate to academic competence beyond the effects of IQ (Blair, 2001; Gottfried, 1990; Masten et al., 2005), but it remains possible that the strength of the relations would be reduced after including IQ.

Despite these limitations, the findings from the current study advance the understanding of the relations between regulatory and social variables and academic competence and provide new information about mechanisms that may explain why children's regulatory abilities are associated with their learning and school success. The results presented here provide evidence that EC is related to academic competence, that the teacher-child relationship and classroom participation partially mediate the relation of EC to



GPA and absences, and that social competence partially mediates the relation of EC to GPA. These models present some possible process mechanisms underlying factors that are associated with children's academic competence, and these findings emphasize the importance of considering regulatory and social influences on academic competence in future models.

## References

- Alexander, K. L., & Entwisle, D. R. (1988). Achievement in the first 2 years of school: Patterns and processes. *Monographs of the Society for Research in Child Development*, 53(2).
- Arnold, D. H., Homrok, S., Ortiz, C., & Stowe, R. M. (1999). Direct observation of peer rejection acts and their temporal relation with aggressive acts. *Early Childhood Research Quarterly*, 14, 183-196.
- Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117, 497-529.
- Berndt, T. J., & Keefe, K. (1995). Friends' influence on adolescents' adjustment to school. *Child Development*, 66, 1312-1329.
- Birch, S. H., & Ladd, G. W. (1997). The teacher-child relationship and children's early school adjustment. *Journal of School Psychology*, 35, 61-79.
- Blair, C. (2001). The early identification of risk for grade retention among African American children at risk for school difficulty. *Applied Developmental Science*, 5, 37-50.
- Blair, C. (2002). School readiness: Integrating cognition and emotion in a neurobiological conceptualization of children's functioning at school entry. *American Psychologist*, 57, 111-127.
- Bronson, M. B., Tivnan, T., & Seppanen, P. S. (1995). Relations between teacher and classroom activity variables and the classroom behaviors of prekindergarten children in Chapter 1 funded programs. *Journal of Applied Developmental Psychology*, 16, 253-282.
- Buhs, E. S., & Ladd, G. W. (2001). Peer rejection as antecedent of young children's school adjustment: An examination of mediating processes. *Developmental Psychology*, 37, 550-560.
- Capaldi, D. M., & Rothbart, M. K. (1992). Development and validation of an early adolescent temperament measure. *Journal of Early Adolescence*, 12, 153-173.
- Caspi, A., Elder, G. H., & Bem, D. J. (1987). Moving against the world: Life-course patterns of explosive children. *Developmental Psychology*, 23, 308-313.
- Children's Action Alliance. (1999). *One in three: Trends, challenges, and opportunities facing Hispanic families in Arizona*. Phoenix, AZ: Author.
- Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, 19, 15-19.
- Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment. *Journal of Family Psychology*, 19, 294-304.
- Dweck, C. S. (1989). Motivation. In A. Lesgold & R. Glaser (Eds.), *Foundations for a psychology of education* (pp. 87-136). Hillsdale, NJ: Erlbaum.
- Eccles, J. S., Wigfield, A., & Schiefele, U. (1998). Motivation to succeed. In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (pp. 1017-1095). New York: Wiley.
- Eisenberg, N. (2006). Introduction. In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (6th ed., pp. 1-23). New York: Wiley.
- Eisenberg, N., & Fabes, R. A. (1998). Prosocial development. In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (pp. 701-778). New York: Wiley.
- Eisenberg, N., Fabes, R. A., Guthrie, I. K., & Reiser, M. (2000). Dispositional emotionality and regulation: Their role in predicting quality of social functioning. *Journal of Personality and Social Psychology*, 78, 136-157.
- Eisenberg, N., Gershoff, E. T., Fabes, R. A., Shepard, S. A., Cumberland, A. J., Losoya, S. H., et al. (2001). Mothers' emotional expressivity and children's behavior problems and social competence: Mediation through children's regulation. *Developmental Psychology*, 37, 475-490.
- Eisenberg, N., Pidada, S., & Liew, J. (2001). The relations of regulation and negative emotionality to Indonesian children's social functioning. *Child Development*, 72, 1747-1763.
- Eisenberg, N., Sadovsky, A., & Spinrad, T. (2005). Associations among emotion-related regulation, language skills, emotion knowledge, and academic outcomes. *New Directions in Child and Adolescent Development*, 109, 109-118.
- Ensminger, M. E., & Slusarcick, A. L. (1992). Paths to high school graduation or dropout: A longitudinal study of a first-grade cohort. *Sociology of Education*, 65, 95-113.
- Fabes, R. A., Martin, C. L., Hanish, L. D., Anders, M. C., & Madden-Derdich, D. A. (2003). Early school competence: The roles of sex-segregated play and effortful control. *Developmental Psychology*, 39, 848-858.
- Finn, J. D. (1993). *School engagement and students at risk* (Publication No. NCES 93470). Washington, DC: U.S Department of Education, National Center of Educational Statistics. (ERIC Document Reproduction Service No. ED 362 322)
- Frome, P. M., & Eccles, J. S. (1998). Parents' influence on children's achievement-related perceptions. *Journal of Personality and Social Psychology*, 74, 435-452.
- Furrer, C., & Skinner, E. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology*, 95, 148-162.
- Gilliom, M., Shaw, D. S., Beck, J. E., Schonberg, M. A., & Lukon, J. L. (2002). Anger regulation in disadvantaged preschool boys: Strategies, antecedents, and the development of self-control. *Developmental Psychology*, 38, 222-235.
- Gottfried, A. E. (1990). Academic intrinsic motivation in young elementary school children. *Journal of Educational Psychology*, 82, 525-538.
- Gottfried, A. E., Fleming, J. S., & Gottfried, A. W. (1994). Role of parental motivational practices in children's academic intrinsic motivation and achievement. *Journal of Educational Psychology*, 86, 104-113.
- Hamre, B. K., & Pianta, R. C. (2001). Early teacher-child relationships and the trajectory of children's school outcomes through eighth grade. *Child Development*, 72, 625-638.
- Harter, S. (1982). The Perceived Competence Scale for Children. *Child Development*, 53, 87-97.
- Haveman, R., & Wolfe, B. (1995). The determinants of children's attainments: A review of methods and findings. *Journal of Economic Literature*, 33, 1829-1878.
- Hill, N. E., & Craft, S. A. (2003). Parent-school involvement and school performance: Mediated pathways among socioeconomically comparable African American and Euro-American families. *Journal of Educational Psychology*, 95, 74-83.
- Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.
- Huffman, L. C., Mehlinger, S. L., & Kerivan, A. S. (2000). *Risk factors for academic and behavioral problems at the beginning of school*. Bethesda, MD: National Institute of Mental Health.
- Hughes, J. N., & Kwok, O.-m. (2006). Classroom engagement mediates the effect of teacher-student support on elementary students' peer acceptance: A prospective analysis. *Journal of School Psychology*, 43, 465-480.
- Jordan, N. C., Kaplan, D., Oláh, L. N., & Locuniak, M. N. (2006). Number



- sense growth in kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child Development*, 77, 153–175.
- Kaufman, P., Alt, M. N., & Chapman, C. D. (2004). *Dropout rates in the United States: 2001* (Publication No. NCES 2005046). Washington, DC: U.S. Government Printing Office.
- Kochanska, G., Coy, K. C., & Murray, K. T. (2001). The development of self-regulation in the first four years of life. *Child Development*, 72, 1091–1111.
- Kochanska, G., Murray, K., & Coy, K. C. (1997). Inhibitory control as a contributor to conscience in childhood: From toddler to early school age. *Child Development*, 68, 263–277.
- Kochanska, G., Murray, K. T., & Harlan, E. T. (2000). Effortful control in early childhood: Continuity and change, antecedents, and implications for social development. *Developmental Psychology*, 36, 220–232.
- Ladd, G. W. (2003). Probing the adaptive significance of children's behavior and relationships in the school context: A child by environment perspective. In R. V. Kail (Ed.), *Advances in child development and behavior* (Vol. 31, pp. 43–104). San Diego, CA: Academic Press.
- Ladd, G. W., Birch, S. H., & Buhs, E. S. (1999). Children's social and scholastic lives in kindergarten: Related spheres of influence? *Child Development*, 70, 1373–1400.
- Ladd, G. W., Buhs, E. S., & Seid, M. (2000). Children's initial sentiments about kindergarten: Is school liking an antecedent of early classroom participation and achievement? *Merrill Palmer Quarterly*, 46, 255–279.
- Ladd, G. W., Kochenderfer, B. J., & Coleman, C. C. (1996). Friendship quality as a predictor of young children's early school adjustment. *Child Development*, 67, 1103–1118.
- Lemerise, E. A., & Arsenio, W. F. (2000). An integrated model of emotion processes and cognition in social information processing. *Child Development*, 71, 107–118.
- Lewit, E. M., & Baker, L. S. (1995). School readiness. *Future of Children*, 5, 128–139.
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7, 83–104.
- Masten, A. S., Roisman, G. I., Long, J. D., Burt, K. B., Obradović, J., Riley, J. R., et al. (2005). Developmental cascades: Linking academic achievement and externalizing and internalizing symptoms over 20 years. *Developmental Psychology*, 41, 733–746.
- McEvoy, A., & Welker, R. (2000). Antisocial behavior, academic failure, and school climate: A critical review. *Journal of Emotional and Behavioral Disorders*, 8, 130–140.
- McGee, R., Partridge, F., Williams, S., & Silva, P. A. (1991). A twelve-year follow-up of preschool hyperactive children. *Journal of the American Academy of Child & Adolescent Psychiatry*, 30, 224–232.
- Meeker, W. Q., Cornwell, L. W., & Aroian, L. A. (1981). *Selected tables in mathematical statistics: Vol. VII. The product of two normally distributed random variables*. Providence, RI: American Mathematical Society.
- Midgley, C., Feldlaufer, H., & Eccles, J. S. (1989). Student/teacher relations and attitudes toward mathematics before and after the transition to junior high school. *Child Development*, 60, 981–992.
- Murray, C., & Greenberg, M. T. (2000). Children's relationship with teachers and bonds with school: An investigation of patterns and correlates in middle childhood. *Journal of School Psychology*, 38, 423–445.
- NICHD Early Child Care Research Network. (2003). Do children's attention processes mediate the link between family predictors and school readiness? *Developmental Psychology*, 39, 581–593.
- NICHD Early Child Care Research Network. (2005). Predicting individual differences in attention, memory, and planning in first graders from experiences at home, child care, and school. *Developmental Psychology*, 41, 99–114.
- Pianta, R. C. (2001). *Student-Teacher Relationship Scale*. Odessa, FL: Psychological Assessment Resources.
- Pianta, R. C., La Paro, K. M., Payne, C., Cox, M. J., & Bradley, R. (2002). The relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *Elementary School Journal*, 102, 225–238.
- Pierce, K. M., Hamm, J. V., & Vandell, D. L. (1999). Experiences in after-school programs and children's adjustment in first-grade classrooms. *Child Development*, 70, 756–767.
- Raver, C. C. (2002). Emotions matter: Making the case for the role of young children's emotional development for early school readiness. *Social Policy Report*, 16, 3–18.
- Resnick, M. D., Bearman, P. S., Blum, R. W., Bauman, K. E., Harris, K. M., Jones, J., et al. (1997). Protecting adolescents from harm: Findings from the National Longitudinal Study on Adolescent Health. *Journal of the American Medical Association*, 278, 823–832.
- Rodriguez, M. L., Mischel, W., & Shoda, Y. (1989). Cognitive person variables in the delay of gratification of older children at risk. *Journal of Personality and Social Psychology*, 57, 358–367.
- Rothbart, M. K., & Bates, J. E. (2006). Temperament. In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (6th ed., pp. 99–166). New York: Wiley.
- Ruff, H. A., & Rothbart, M. K. (1996). *Attention in early development: Themes and variations*. London: Oxford University Press.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177.
- Schultz, D., Izard, C. E., & Ackerman, B. P. (2000). Children's anger attribution bias: Relations to family environment and social adjustment. *Social Development*, 9, 284–301.
- Shoda, Y., Mischel, W., & Peake, P. K. (1990). Predicting adolescent cognitive and self-regulatory competencies from preschool delay of gratification: Identifying diagnostic conditions. *Developmental Psychology*, 26, 978–986.
- Shores, R. E., & Wehby, J. H. (1999). Analyzing the classroom social behavior of students with EBD. *Journal of Emotional and Behavioral Disorders*, 7, 194–199.
- Simpkins, S. D., Davis-Kean, P. E., & Eccles, J. S. (2006). Math and science motivation: A longitudinal examination of the links between choices and beliefs. *Developmental Psychology*, 42, 70–83.
- Stoops, N. (2004). *Educational attainment in the United States: 2003*. Washington, DC: U.S. Census Bureau.
- Tamis-LeMonda, C. S., & Bornstein, M. H. (1989). Habituation and maternal encouragement of attention in infancy as predictors of toddler language, play, and representational competence. *Child Development*, 60, 738–751.
- U.S. Bureau of the Census. (2004). *U.S. interim projections by age, sex, race, and Hispanic origin*. Washington, DC: U.S. Government Printing Office.
- U.S. Department of Education. (2004). *Stronger accountability: Reaching out, raising Hispanic achievement*. Washington, DC: Author.
- Valiente, C., Eisenberg, N., Fabes, R. A., Shepard, S. A., Cumberland, A. J., & Losoya, S. H. (2004). Prediction of children's empathy-related responding from their effortful control and parents' expressivity. *Developmental Psychology*, 40, 911–926.
- Valiente, C., Eisenberg, N., Smith, C. L., Reiser, M., Fabes, R. A., Losoya, S., et al. (2003). The relations of effortful control and reactive control to children's externalizing problems: A longitudinal assessment. *Journal of Personality*, 71, 1171–1196.
- Valiente, C., Lemery-Chalfant, K. S., & Castro, K. S. (2007). Children's effortful control and academic competence: Mediation through school liking. *Merrill-Palmer Quarterly*, 53, 1–25.
- Welsh, M., Parke, R. D., Widaman, K., & O'Neil, R. (2001). Linkages

- between children's social and academic competence: A longitudinal analysis. *Journal of School Psychology*, 39, 463-482.
- Wentzel, K. R. (1999). Social-motivational processes and interpersonal relationships: Implications for understanding motivation at school. *Journal of Educational Psychology*, 91, 76-97.
- West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56-75). Thousand Oaks, CA: Sage.
- Wigfield, A., Eccles, J. S., Schiefele, U., Roeser, R., & Davis-Kean, P. (2006). Development of achievement motivation. In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (6th ed., pp. 933-1002). New York: Wiley.
- Willson, V. L., & Hughes, J. N. (2006). Retention of Hispanic/Latino

- students in first grade: Child, parent, teacher, school, and peer predictors. *Journal of School Psychology*, 44, 31-49.
- Zhou, Q., Eisenberg, N., Wang, Y., & Reiser, M. (2004). Chinese children's effortful control and dispositional anger/frustration: Relations to parenting styles and children's social functioning. *Developmental Psychology*, 40, 352-366.
- Zimmerman, B. J. (1998). Developing self-fulfilling cycles of academic regulation: An analysis of exemplary instructional models. In D. H. Schunk (Ed.), *Self-regulated learning: From teaching to self-reflective practice* (pp. 1-19). New York: Guilford Press.

Received October 31, 2006  
 Revision received May 31, 2007  
 Accepted June 12, 2007 ■

## ORDER FORM

Start my 2008 subscription to the *Journal of Educational Psychology* ISSN: 0022-0663

\_\_\_\_\_ \$73.00, APA MEMBER/AFFILIATE \_\_\_\_\_  
 \_\_\_\_\_ \$161.00, INDIVIDUAL NONMEMBER \_\_\_\_\_  
 \_\_\_\_\_ \$450.00, INSTITUTION \_\_\_\_\_  
 In DC add 5.75% / In MD add 6% sales tax \_\_\_\_\_  
**TOTAL AMOUNT ENCLOSED** \$ \_\_\_\_\_

**Subscription orders must be prepaid.** (Subscriptions are on a calendar year basis only.) Allow 4-6 weeks for delivery of the first issue. Call for international subscription rates.



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

**SEND THIS ORDER FORM TO:**  
 American Psychological Association  
 Subscriptions  
 750 First Street, NE  
 Washington, DC 20002-4242

Or call **800-374-2721**, fax **202-336-5568**.  
 TDD/TTY **202-336-6123**.  
 For subscription information, e-mail:  
**subscriptions@apa.org**

☐ **Check enclosed** (make payable to APA)

**Charge my:** ☐ VISA ☐ MasterCard ☐ American Express

Cardholder Name \_\_\_\_\_

Card No. \_\_\_\_\_ Exp. Date \_\_\_\_\_

\_\_\_\_\_  
 Signature (Required for Charge)

### BILLING ADDRESS:

Street \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

Daytime Phone \_\_\_\_\_

E-mail \_\_\_\_\_

### MAIL TO:

Name \_\_\_\_\_

Address \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

APA Member # \_\_\_\_\_ EDUA08



# A Multilevel Perspective on Gender in Classroom Motivation and Climate: Potential Benefits of Male Teachers for Boys?

Herbert W. Marsh  
University of Oxford

Andrew J. Martin  
University of Sydney

Jacqueline H. S. Cheng  
University of Oxford

Are boys better motivated by male than female teachers in high school math, science, and English classes, and can these differences be explained by classroom climate? Using a cross-classified multilevel model with 5 levels (school, teacher, class, student, subject), the authors found little or no support for this contention. In general (except in terms of anxiety and persistence), girls were better motivated than boys, and these differences tended to generalize over student age and school subject in classes taught by both male and female teachers. Student perceptions of classroom climate were more specific to the group of students within a particular class than to the teacher who taught the class and had moderate to large effects on the motivation of individual students. The surprisingly small amounts of variance explained in motivation by student gender and age, teacher gender, school subject, and their interactions support a gender invariance and similarities model but not theoretical predictions based on gender stereotype, gender intensification, and gender matching perspectives.

**Keywords:** student and teacher gender, student motivation, classroom climate, multilevel modeling with cross-classification

Do boys fare better in classes taught by male teachers? In a recent article in the *Washington Post* newspaper, columnist Ben Feller (2006) stated, "For all the differences between the sexes, here's one that might stir up debate in the teacher's lounge: Boys learn more from men and girls learn more from women." He then went on to describe research in support of this claim by Dee (2006) that has attracted media publicity across the United States. Included in his article was an array of quotes in support for and against the basic contention that boys are benefited by male teachers. Clearly, the debate on this topic is not limited to the United States, as an Australian Labor Party (the major opposition political party at the time) policy document leading up to a recent federal election stated, "Now, more than ever, young boys need contact with men who can offer positive role models and mentor them in the right direction (p. 1). . . . Labor wants to see many more male teachers teaching and making a difference to the lives

of young boys in our schools" (Australian Labor Party, 2004, p. 4). In contrast, on the basis of interviews with teachers in Iceland, Jóhannesson (2004) was especially critical of the "myth that boys in particular need 'male role models'" (p. 33). Weaver-Hightower (2003) reviewed the political, research, and practice-oriented causes of the recent upsurge in interest in boys' education and what he called the "boy turn," which has focused on the schooling experiences and outcomes of boys.

Following from this controversy, which bridges popular media, academic research, and policy development all over the world, the focus of this study is the question, "Are boys better motivated by male teachers?" In addressing this issue it is, of course, critical to evaluate motivation for both male and female students who are taught by male and female teachers. Hence, whereas our focus is on motivating boys as the starting point for our research, the results have implications for girls as well. In pursuing an answer to our question, we begin by briefly reviewing theoretical and empirical research on gender differences, consider how teacher gender and student gender might interact (the matching hypothesis), and apply new statistical procedures—cross-classified multilevel modeling (MLM)—that are ideally suited to exploring interactions between student gender and teacher gender.

Addressing teacher effects on student motivation also provides an opportunity to explore other important and cognate issues, particularly under a multilevel framework such as that developed here. Thus, for example, we argue that to the extent that teachers affect the motivation levels of their students, these effects are likely to be mediated at least in part by classroom climate. Hence, while addressing teacher effects, we are also able to assess—using appropriate analytic methods—the role of class climate in student

---

Herbert W. Marsh and Jacqueline H. S. Cheng, Department of Education, University of Oxford, Oxford, England; Andrew J. Martin, Faculty of Education and Social Work, University of Sydney, Sydney, Australia.

Work on the present investigation was conducted, in part, during the periods when each of us was part of the SELF Research Centre at the University of Western Sydney and while Andrew J. Martin was Visiting Research Fellow at the University of Oxford. This research was supported in part by a grant from the Australian Research Council.

Correspondence concerning this article should be addressed to Herbert W. Marsh, Department of Education, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY, England, or to Andrew J. Martin, Faculty of Education and Social Work, University of Sydney, Sydney NSW 2006, Australia. E-mail: herb.marsh@education.ox.ac.uk or a.martin@edfac.usyd.edu.au

motivation and attempt to disentangle the respective roles of teacher and class composition in shaping student motivation and classroom climates, which are posited to be one basis of motivation in individual students. In this respect, the present investigation is a methodological–substantive synergy (Marsh & Hau, 2007) involving new and evolving statistical analyses—in this case, MLM—that allow us to evaluate substantive issues that could not easily or appropriately be pursued without such advances. Hence, this study addresses important substantive issues that are not effectively evaluated with existing statistical procedures, thereby providing a catalyst for the development of better methodology.

### The Gender Invariance Hypothesis and the Matching Hypothesis

In education research and policy—indeed, in the psychological and social science disciplines more generally—there is a preoccupation with gender differences (see Hyde, 2005). Fueled in part by a null hypothesis testing perspective, given a sufficiently large sample size, there will be statistically significant gender differences for most psychosocial variables. However, this focus on gender differences tends to ignore the strong support for gender similarities and the typically small sizes of gender differences. As emphasized by Hyde (2005) in her meta-analytic review of meta-analyses of gender differences, men and women are more similar than different on most variables. She argued that overinflated interpretations of small gender differences can do much harm and lead to counterproductive policies. In education in particular, there is a long history of research about the effect of the gender of both students and teachers. In this article, we test a matching hypothesis, coupled with gender stereotype and gender intensification models, about how gender differences in student motivation in different school subjects are influenced by the gender of the teacher.

Marsh (1989a, 1989b) reviewed research into gender differences, particularly that by Eccles and colleagues (e.g., Eccles, 1987; also see Eccles & Wigfield, 2002), and into differential socialization processes proposed by many other researchers (e.g., Maccoby & Jacklin, 1974). He posited and tested a differential socialization hypothesis in which “sex-linked differences in socialization patterns may fail to reinforce adequately boys’ positive attitudes, expectations, and performance in verbal areas as well as failing to reinforce adequately girls’ positive attitudes, expectations, self-concepts, and performance in mathematics” (Marsh, 1989b, p. 195). Although he found small gender-stereotypic differences for math and verbal constructs that were consistent with other research (e.g., Hyde, 2005), he also identified a more long-term perspective based on nationally representative samples showing that gender differences favoring girls were becoming larger, whereas gender differences favoring boys were becoming smaller (also see Martin & Marsh, 2005). Indeed, this trend of increasingly poorer academic outcomes for boys has fueled, in part, the question of whether boys are benefited by having male teachers. However, Marsh (1989b) found that other aspects of the gender-stereotypic model were not supported—relations among variables were largely invariant over gender.

Marsh (1993) provided an alternative test of the differential socialization model, in which it was predicted that English self-concept would be more highly related to academic and general

self-concept for girls, math self-concept would be more highly related to academic and general self-concept for boys, and these gender differences would grow larger with age. These predictions were consistent with Eccles’s (1987) finding that gender differences in the value placed on math and verbal competence grew larger with age and with what Hill and Lynch (1983) called gender-role intensification, in which conformity to gender-role stereotypes becomes increasingly important with age. However, Marsh (1993) found no support for these predictions, as relations were similar across eight groups (2 gender  $\times$  4 adolescent age groups). On the basis of his research, Marsh (1993; Marsh, Hau, Sung, & Yu, 2007; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005; Martin & Marsh, 2005) argued for a gender-invariant model in which relations among gender-stereotypic constructs considered in his research were similar for boys and girls, a conclusion consistent with the gender similarities hypothesis posited by Hyde (2005).

In the present article, we extend this previous research to include a matching hypothesis in which student gender and teacher gender are hypothesized to interact—favorably when student and teacher genders match, but unfavorably when they do not. Previous research has sometimes tested a matching hypothesis as a potential source of bias. Thus, for example, Jayasinghe, Marsh, and Bond (2001, 2002) tested a matching hypothesis that peer reviews of grant applications would be higher when individual characteristics of the external assessor (e.g., gender, age, academic rank, status and type of institution) matched those of the applicant. However, in their MLM analysis covering all major science disciplines, they found no support for a matching hypothesis—operationalized in terms of statistically significant interactions between characteristics of the assessor and the applicant. In particular, there was no evidence that female (or male) reviewers gave systematically higher or lower ratings to proposals authored by female (or male) researchers.

In another application of the matching hypothesis, Marsh (1987b, 2007b) evaluated whether university students gave higher ratings of teaching effectiveness to same-gender teachers than to opposite-gender teachers. Support for such a matching hypothesis might imply a form of gender bias in the teacher evaluations or, alternatively, the claim that men are more effective teachers for male students, whereas women are more effective teachers of female students. However, Marsh argued that there was little or no support for a matching hypothesis—and, by implication, either of these alternative interpretations (also see Feldman, 1992, 1993). Although Centra and Gaubatz (2000) found some support for a matching hypothesis on the basis of students’ evaluations of university teaching, he emphasized that the sizes of the effects were too small to be substantively important.

In the present investigation, we combine theoretical perspectives from gender-stereotypic, gender intensification, and matching perspectives with new developments in MLM to test whether boys are better motivated by male teachers. Inherent in any discussion of the teacher effects on the motivation levels of individual students within a classroom is the motivational orientations within the classroom—the classroom climate. Hence, teacher effects on student motivation—and their interaction with other student characteristics, such as student gender—are likely to be mediated, at least in part, by classroom climate. Here we bring together the study of individual student motivation, student and teacher gender,



and classroom climate within an appropriate MLM framework that is ideally suited to this purpose and demonstrate the new applications of this important statistical procedure, which we now discuss in greater detail.

### The Hierarchical Nature of Motivation and Classroom Climate: An MLM Approach

Our intent is to apply important advances in the application of MLM (also referred to as *hierarchical linear modeling*) to the study of motivation, teacher effects, and classroom motivation climates (also see Ntoumanis & Vazou, 2005; Papaioannou, Marsh, & Theodorakis, 2004; Raudenbush, Rowan, & Kang, 1991). Particularly in education—and the social sciences more generally—data typically have a multilevel structure in which individuals are clustered into groups (e.g., classes), which might be clustered into higher level administrative units (e.g., schools). Individual student characteristics and those of the classes to which students belong are usually confounded because individuals are not randomly assigned. Students are typically more similar to other students within their class than they are to other students from different classes.

When data have a multilevel structure, single-level analyses that ignore this multilevel structure are typically invalid (violating statistical assumptions in a way that increases the likelihood of a Type I error; for further discussion, see Goldstein, 2003; Raudenbush & Bryk, 2002). Although these statistical issues dictate that MLM analyses should be used routinely, a substantive MLM focus opens up important new perspectives about how constructs operate at the individual and group levels. However, an MLM perspective also provides new challenges for theory, research, measurement, and practice. Thus, for example, the same construct measured at the level of the individual and at the level of the group may lead to fundamentally different interpretations (e.g., Marsh & Hau, 2003). Furthermore, measures that have good psychometric properties at the individual level may not be satisfactory at the group level (e.g., Marsh, Rowe, & Martin, 2002). Hence, measures demonstrated to be effective at one level must also be evaluated at other levels to which they are applied via a traditional construct validity approach. In addition, there may be substantively important questions of practical significance that involve interactions between individual- and group-level variables. For example, in the present investigation we posit a matching hypothesis in which boys are better motivated by male teachers. An evaluation of interactions between variables at the individual student level and the teacher level is inherently an MLM issue. Hence, the MLM method provides a much richer and more appropriate methodological approach to evaluating motivational climate than would be possible with traditional single-level approaches, which ignore the clustering of individuals within groups.

Particularly when the focus of the research is on an inherently class-level variable, such as climate, it makes no sense to ignore the class level in the analysis. Traditionally, researchers have distinguished between the motivational levels of individual students and the perceptions of the individual students about the classroom motivational climate. Although the motivational levels of individual students are clearly designed to be an individual-level variable, the appropriate level of analysis for classroom climate is not straightforward. Historically, individual student perceptions of

classroom climate have typically been analyzed with single-level analyses that treat classroom climate as an individual student-level variable (e.g., Ames, 1992; Ames & Archer, 1988; Levesque, Zuehlke, Stanek, & Ryan, 2004; Papaioannou, 1995; Seifriz, Duda, & Chi, 1992). In contrast, we argue that classroom climate is inherently a class-level construct. When climate measures are based on responses about the overall class climate by a single person for each class (e.g., teacher or external observer), then it is clear that the level of analysis is the class. When each student in a class evaluates the classroom climate, it is possible to consider responses at both the individual and the classroom levels. Even here, however, we argue that the most appropriate measures of the classroom climate are aggregates of individual student perceptions. It is, however, important to evaluate classroom perceptions of individual students as well. If, for example, there is little or no agreement about classroom climate among students within the same class, then support for the construct validity of the perceived climate ratings is dubious (in the same way as lack of agreement among items designed to measure the same construct undermines support for the construct validity of the construct). This is not to say, of course, that the perceptions of each student and how these individual student perceptions differ from the aggregated perceptions are not also important. Fortunately, recent advances in MLM provide new opportunities to combine constructs from different levels of analysis. Hence, the issue is not whether research should be done at the level of the individual participant or the group level but how best to combine measures from multiple levels of measurement.

In classroom climate research, classroom climate is often posited to be a function of the teacher who teaches a class. However, as is the case with all teacher effects, a viable alternative is that classroom climate is a function, at least in part, of the students who compose the class. Typically, the effects of a teacher and the class are completely confounded and cannot be distinguished. However, when the design of an appropriate MLM study contains more than one class per teacher, it is possible to distinguish between these alternative explanations, evaluating the extent to which effects of a given teacher generalize over different classes taught by the same teacher. This distinction is particularly important in studies such as ours, in which we are specifically concerned with the generalizability of effects associated with characteristics of different teachers (i.e., teacher gender).

### Substantive Hypotheses and Research Questions

The present study encompasses substantive questions regarding the effects of teacher gender on student motivation and the differential levels at which variation in motivation occurs (e.g., student, class, teacher, and school levels). However, these issues also enable new methodological questions to be explored, such as the appropriate means by which to disentangle teacher and class composition effects and, in particular, the role of cross-classification analysis to achieve this. Hence, the present study, using cross-classification MLM, explores (a) the extent to which student gender, teacher gender, and their interaction impact on dependent variables consisting of multiple dimensions of student motivation, affective outcomes, and corresponding student perceptions of classroom climate and (b) the extent to which motivation, outcomes, and classroom climate perceptions vary as a function of



school subject, student, class, teacher, and school levels. To provide a substantive focus for our MLM analyses, we offer the following hypotheses and research questions adapted for the purposes of the present investigation from gender-stereotypic and gender intensification models and from the matching hypothesis. Consistent with concerns about boys increasingly falling behind girls in a variety of educational outcomes, which is an underlying theme driving the present investigation, we expected girls to be generally better motivated than boys (i.e., higher for adaptive motivations and outcomes, lower for maladaptive motivations). However, our focus is on how such gender differences vary as a function of the gender of the teacher, school subject, age, and that particular component of motivation. On the basis of our review, we propose the following research hypotheses and questions.

### *Gender-Stereotypic Differences and Gender Intensification*

Motivation, outcomes, and perceptions of classroom climate will be more favorable (i.e., higher for adaptive motivations and outcomes, lower for maladaptive motivations) in gender-stereotypic subjects (i.e., more favorable for girls in English and for boys in math and science). Gender-stereotypic differences in motivation will grow larger with age as students move into adolescence.

### *Matching Hypothesis*

Boys, in particular, will be more motivated by male teachers than by female teachers, but girls will also be more motivated by female teachers. In support of the gender intensification hypothesis, this Student Gender  $\times$  Teacher Gender interaction is predicted to become stronger with age.

We leave as a research question whether support for the matching hypothesis varies as a function of the school subject. It could be argued that effects of student-teacher gender matching are particularly large in subjects that go against traditional gender stereotypes (positive effects of male English teachers for boys and of female math and science teachers for girls), where motivation is more likely to be lower. Alternatively, it could be argued that male teachers are particularly able to motivate boys in math and sciences, whereas female teachers are particularly able to motivate girls in English classes.

### *Gender Invariance and Similarities Model*

In contrast to predictions based on gender stereotypes and the matching hypothesis, this model predicts that gender differences in motivation and classroom climate tend to be small (gender similarities) and do not vary much with school subject, age, or gender of the teacher (gender invariance). Support for the gender similarities prediction comes from Hyde (2005), whereas empirical and theoretical support for the invariance hypothesis is from Marsh (1993; Marsh et al., 2005, 2007; Martin & Marsh, 2005). However, it is important to emphasize that there is apparently very little rigorous empirical support for either the matching hypothesis (that boys are better motivated by men and girls are better motivated by women) or the invariance hypothesis (that teacher gender and student gender do not interact). Nevertheless, we contend that

there is sufficient justification for the invariance hypothesis to warrant positing it as a specific hypothesis to be tested. In particular, although it could be argued that the invariance hypothesis is implicit in the matching hypothesis (i.e., it is the null hypothesis in tests of Student Gender  $\times$  Teacher Gender interactions), we chose to make it explicit as a viable alternative hypothesis. Whereas arguing for the null hypothesis is always a precarious undertaking, the juxtaposition of competing sets of hypotheses involves a critical evaluation of the practical implications of the direction and sizes of effects.

A second important component of our research is classroom climate at the levels of student, class, and teacher and how it affects student motivation. In part, our research builds on earlier research by Papaioannou et al. (2004) that evaluated classroom motivational climates in physical education classes. They found that teacher effects were very small for individual student motivation but, not surprisingly, were larger for classroom motivational climate. Furthermore, the motivational climates in these physical education classes were more a function of the teachers who taught the class than of the particular group of students who were in the class—motivation climates associated with a particular teacher tended to generalize across different classes taught by the same teacher. Here we evaluate how these results generalize to mathematics, English, and science classes and the implications of these results to gender differences associated with student and teacher gender.

### *Motivation and Climate at the Level of the Student, Class, and Teacher*

There are distinguishable effects in classroom climates associated with the teacher who teaches a class (teacher-level effects) and the particular group of students in a specific class (class-level effects). While we expect that there are teacher-level effects that generalize across the different classes teachers teach, there may also be class-level effects that are idiosyncratic to particular groups of students and that do not generalize to other classes taught by the same teacher (e.g., Papaioannou et al., 2004). There is a strong body of research at the university level showing that students' evaluations of effective teaching are primarily a function of the teacher who teaches a class rather than of the class or group of students who rate a particular class (Marsh, 1987b, 2007b). However, because there is little research on this issue at the primary or secondary level that focuses on motivation and classroom climate, we posit this as a research question to be explored.

Effects of teachers and classrooms are stronger for student perceptions of classroom climate than for the corresponding measures of individual student motivation and outcomes (e.g., Papaioannou et al., 2004). Conversely, it follows that individual student effects should be stronger for student motivation than for classroom climate perceptions (e.g., Papaioannou et al., 2004). Taken together, these two predictions imply that perceptions of classroom climate should be more specific to particular classes and teachers, whereas motivation levels of individual students should be more specific to individual students. This set of hypotheses merely means that there should be better agreement among students within the same class when they rate the class in terms of classroom climate than when they rate their own individual levels of motivation. Indeed, failure to support this commonsense hy-



pothesis would call into question the construct validity of the classroom climate ratings. The critical aspect of this prediction is that there should be substantial agreement among students within the same class in terms of their classroom climate ratings. To better understand this prediction, it may be useful to consider a somewhat analogous situation for items (analogous to different students within the same class) and factors (analogous to the different classes). The analogous prediction from the item-factor perspective is that items designed to measure the same factor should be more highly correlated with each other than with items designed to measure a different factor. Similarly, students who are making ratings of the same classroom should show greater agreement among each other than with other students who are rating a different classroom.

Individual motivation levels for the same student do generalize over different school subjects. Research by Marsh, Martin, and Debus (2001) suggests that for motivation constructs such as those measured here, there is substantial agreement in motivation levels across math and verbal subjects, although there are some self-belief constructs in which the relation is close to zero (i.e., math and verbal self-concepts; Marsh, 2007a). Hence, we expect that the student effect—the extent to which motivation varies across different school subjects—will be substantial for individual ratings of motivation but that its size should vary across different motivation constructs.

To what extent does classroom climate—at the level of the classroom—have a substantial influence on individual student motivation? To what extent are classroom and teacher effects on individual student motivation mediated by classroom climate? While these are critical research questions, there is not sufficient research from an MLM perspective (which we argue is necessary to address this issue) to make a priori predictions. However, an affirmative response to these questions would provide support for the construct validity of climate ratings.

## Method

### *Sample and Procedure*

The present study focuses on Year 8 and 10 high school students in their mathematics, English, and science classes. The sample comprises 964 high school students from five Australian coeducational government schools: 60% in Year 8 (junior high school, mostly 12 and 13 years of age), and 40% in Year 10 (middle high school, mostly 15 and 16 years of age). Nearly half (48%) the respondents were girls, and 52% were boys. The mean age was 14.30 ( $SD = 1.12$ ) years. In total, 101 classrooms, taught by a total of 62 teachers (58% female), were surveyed.

Teachers administered the Motivation and Engagement Scale—High School (MES-HS; Martin, 2001, 2003d, 2006a, 2007) as well as affective outcomes of relevance. In addition, they completed single-item classroom climate scales designed to parallel the motivation and outcome scales. The rating scale was first explained, and a sample item was presented. Students were then asked to complete the instrument on their own and to return the completed instrument to the teacher at the end of class. It is important to note that students rated their motivation and engagement in the corresponding class (i.e., mathematics motivation was evaluated in math classes, English motivation in English classes,

and science motivation in science classes). Students completed the math survey in a mathematics class ( $n = 964$ ), and most completed either an English survey in their English class ( $n = 331$ ) or a science survey in their science class ( $n = 406$ ).

### *Materials*

Materials considered here were based on responses to the MES-HS (Martin, 2001, 2003d, 2006b, 2007), other affective outcomes constructs deemed to be of relevance to the breadth of students' experience in the classroom, and corresponding measures of classroom climate (see Martin, 2007, for a full description of the development and origins of MES-HS scales). The MES-HS assesses six adaptive motivation scales (three cognitive and three behavioral) and five maladaptive motivation scales (three cognitive and two behavioral). Each of the 11 factors comprises 4 items—44 items in all. To each item, students rate themselves on a scale of 1 (*strongly disagree*) to 7 (*strongly agree*). Previous results (Green, Martin, & Marsh, 2006; Martin & Marsh, 2005; also see Martin, 2001, 2003d, 2007; Martin & Marsh, 2006) provided strong psychometric support for the factor structure of the instrument. A 16-factor confirmatory factor analysis based on the 44-item (11-factor) MES-HS items and the additional 20 items assessing each of the five additional affective outcomes yielded an excellent fit to the data (nonnormed fit index = .97; comparative fit index = .97; root-mean-square error of approximation = .047). Factor loadings relating each item to its a priori factor were consistently substantial, and reliability (coefficient alpha) estimates for the 16 scales (varying from .74 to .89) were good given the brevity of the 4-item factors. Martin (2001, 2003d, 2007) also showed that the MES-HS has a sound factor structure; comprises reliable and approximately normally distributed dimensions; and is significantly associated with literacy, numeracy, and achievement in mathematics and English.

*Adaptive dimensions of motivation and engagement.* Each adaptive dimension fell into one of two groups: cognitions and behaviors. Adaptive cognitions included self-efficacy, mastery orientation, and valuing. Adaptive behaviors included persistence, planning, and task management. In terms of adaptive cognitions, *self-efficacy* (e.g., "If I try hard, I believe I can do my schoolwork well") is students' belief and confidence in their ability to understand or to do well in their schoolwork or studies, to meet challenges they face, and to perform to the best of their ability. *Valuing* (e.g., "Learning at school is important") is how much students believe what they do and learn at school is useful, important, and relevant to them or to the world in general. *Mastery orientation* (e.g., "I feel very pleased with myself when I really understand what I'm taught at school") entails being focused on learning for the development of mastery, competence, and success by dint of and as reflective of effort and skill development.

In terms of adaptive behaviors, *planning* (e.g., "Before I start an assignment, I plan out how I am going to do it") refers to how much students plan their schoolwork and projects and how much they keep track of their progress as they are doing that work. *Task management* (e.g., "When I study, I usually study in places where I can concentrate") refers to the way students use their time, organize themselves, and choose and arrange where they do their schoolwork and study. *Persistence* (e.g., "If I can't understand my schoolwork at first, I keep going over it until I do") is how much



students keep trying to work out an answer or to understand a problem even when that problem is difficult or challenging.

*Maladaptive dimensions of motivation and engagement.* Each maladaptive dimension fell into one of two groups: cognitions and behaviors. Maladaptive cognitive dimensions were anxiety, failure avoidance, and uncertain control. *Anxiety* (e.g., "When exams and assignments are coming up, I worry a lot") had two parts: feeling nervous and worrying. Feeling nervous is the uneasy or sick feeling students get when they think about their schoolwork, assignments, or tests. Worrying is their fear of not doing very well in their schoolwork, assignments, or tests. *Failure avoidance* (e.g., "Often the main reason I work at school is because I don't want to disappoint my parents") reflects an orientation whereby the main reason students try at school is to avoid doing poorly, being seen to do poorly, or the negative consequences of doing poorly. *Uncertain control* (e.g., "When I get a bad mark I'm often unsure how I'm going to avoid getting that mark again") assesses students' uncertainty about how to do well or how to avoid doing poorly.

Maladaptive behavioral dimensions were self-handicapping and disengagement. *Self-handicapping* (e.g., "I sometimes don't study very hard before exams so I have an excuse if I don't do as well as I hoped") assesses an orientation whereby students do things that reduce their chances of success at school in an attempt to deflect the cause of possible poor performance away from a perceived lack of ability or competence. Examples are putting off doing a project or wasting time while they are meant to be doing their schoolwork or preparing for an upcoming assignment or test. *Disengagement* (e.g., "I've pretty much given up being involved in things at school") reflects an orientation whereby students feel like giving up in particular school activities or at school more generally. Students high in disengagement tend to accept failure and behave in ways that reflect helplessness.

*Affective outcomes.* To conduct a more expansive analysis of the issues under focus, we were also interested in exploring the nature of effects on some other conceptually relevant educational constructs. To this end, the sample was also administered multi-item scales that explored their enjoyment of the subject (e.g., "I enjoy this subject"), class participation (e.g., "I get involved in things we do in class"), educational aspirations (e.g., "I'd like to continue studying or training in this subject after I complete school"), teacher-student relationships (e.g., "I get along well with my teacher"), and academic resilience (e.g., "I think I'm good at dealing with schoolwork pressures"). Psychometric properties of these scales are presented below.

*Classroom climate.* For each of the 16 motivation and affective outcome scales, students also made a parallel rating on a single-item scale designed to assess classroom climate. Each of these items began with the phrase "In this class, students . . ." (e.g., "In this class, students believe that they can do a good job on their schoolwork," "In this class, students enjoy this subject," "In this class, students are quite anxious about schoolwork and tests," and "In this class, students have a good relationship with the teacher"), and students responded on the same 7-point response scale as for the other items. It is important to note that while ratings of the 16 motivation and outcome scales were in relation to the individual student completing the survey, the classroom climate ratings were generalized ratings of perceptions about students from the same class as a whole.

### Statistical Analysis: MLM

*Cross-classification.* In traditional MLM, units have a purely hierarchical or nested structure in which units at a lower level (e.g., students) occur at only one level of a higher order unit (e.g., schools). In cross-classified analyses, however, a unit may be classified along more than one dimension (Browne, 2005; Goldstein, 2003; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). Typical applications include, for example, students classified by their school and neighborhood when students from the same neighborhood attend different schools (i.e., neighborhoods are not nested under schools). Failure to account for such cross-classification can substantially bias estimates of variance explained by the different levels. In the present study, classes were nested under teachers (each class was taught by only one teacher), and teachers were nested under schools. Although each combination of student and subject was nested under class, student and subject were cross-classified. Because students had data for different subjects, the data structure was a cross-classified structure rather than a multiple membership model. While cross-classified analyses have not been handled easily by MLM statistical packages until recently, they are facilitated by recent developments in MLMs (see Browne, 2005, for further discussion of the Markov chain Monte Carlo estimation procedures implemented in Version 2 of MLwiN), providing a much stronger statistical basis for testing substantive research hypotheses and questions raised in the present investigation.

*MLM design of the present investigation.* For purposes of the present investigation, the data were conceptualized as a five-level cross-classified MLM (levels were school, teacher, class, student, and subject). The MLM analyses were conducted with MLwiN Version 2.02, with Markov chain Monte Carlo estimation used for cross-classified models (Browne, 2005; also see Rasbash, Steele, Browne, & Prosser, 2004). Classes were nested under teachers, teachers were nested under schools, and each combination of student and subject was nested under class. However, student and subject were cross-classified. In preliminary analyses, a baseline variance components model (Rasbash et al., 2004) or intercept-only model (Hox, 1995) was used to evaluate how much variation in each of the dependent measures (motivation, affective outcomes, classroom climate perceptions) could be attributed to each level. Following variance components models, the major focus of analyses was on a set of MLM path models to test the effects of student and teacher gender, subject (math, English, and science), grade (8th and 10th grade), and their interactions on various dependent variables.

Several data transformations were used to facilitate interpretations and infer interaction effects. We standardized ( $z$  scored) all independent and dependent variables to have  $M = 0$  and  $SD = 1$  across the entire sample (Marsh & Rowe, 1996; also see Aiken & West, 1991; Raudenbush & Bryk, 2002). Product terms were used to test interaction effects. In constructing these interaction effects, we used the product of individual ( $z$  score) standardized variables, but product terms were not restandardized (i.e., were left in the metric of variables composing the product terms).

In MLM studies, the individual is typically taken to be the lowest level (e.g., students nested within classes, classes nested within teachers, teachers nested within schools). In this case, variance associated with individuals is part of the residual variance



term, so that variance specifically associated with differences between individuals cannot be differentiated from other, idiosyncratic sources of variance that contribute to the residual variance term. However, it is possible to introduce additional levels that are nested under individuals. Thus, for example, in longitudinal studies, occasions are nested under individuals, so that individual-level variance indicates how well responses by the same individual generalize over time. In the present investigation, we introduced school subject (math, English, science) as an additional level and evaluated the extent to which motivational levels and classroom climate perceptions of the same student generalized across different school subjects. In particular, we posited that individual student motivation should generalize over subjects to a much greater extent than for the corresponding measures of classroom climate.

Another interesting feature of our MLM design is that many teachers taught more than one class within the same subject area (math, English, and science). Typically, when there are data for only one class for each teacher, the effects of the teacher and the group of students within this class are completely confounded; there is no basis for determining the extent to which teacher effects generalize over different classes taught by the same teacher. In the present investigation, however, class—the group of students within the class—was nested under teacher, so that it is possible to determine the extent to which effects were specific to one group of

students or generalized across different classes taught by the same teacher.

Results

Preliminary (Single-Level) Analyses

We began with preliminary, single-level analyses to provide a descriptive (nonstatistical) overview of the results before moving to the more complicated cross-classified MLM analyses. Because these single-level analyses ignore the complicated multilevel structure of the data, tests of statistical significance for these results are biased and so are not discussed. In Figure 1 we present box-plot graphs for each of the 11 motivation scales as a function of student gender, teacher gender, and school subject—keeping in mind that the first 6 (of 11) scales were positively oriented, adaptive motivational constructs, whereas the last 5 were negatively oriented, maladaptive motivational constructs. For all 6 positively oriented motivations, girls tended to have higher motivations than boys, and this trend appeared to generalize over classes taught by male and female teachers. For negatively oriented motivations, gender differences appeared to be small—except for anxiety, for which girls were more anxious than boys. Again, however, these gender differences (or lack thereof) tended to be similar for male and female

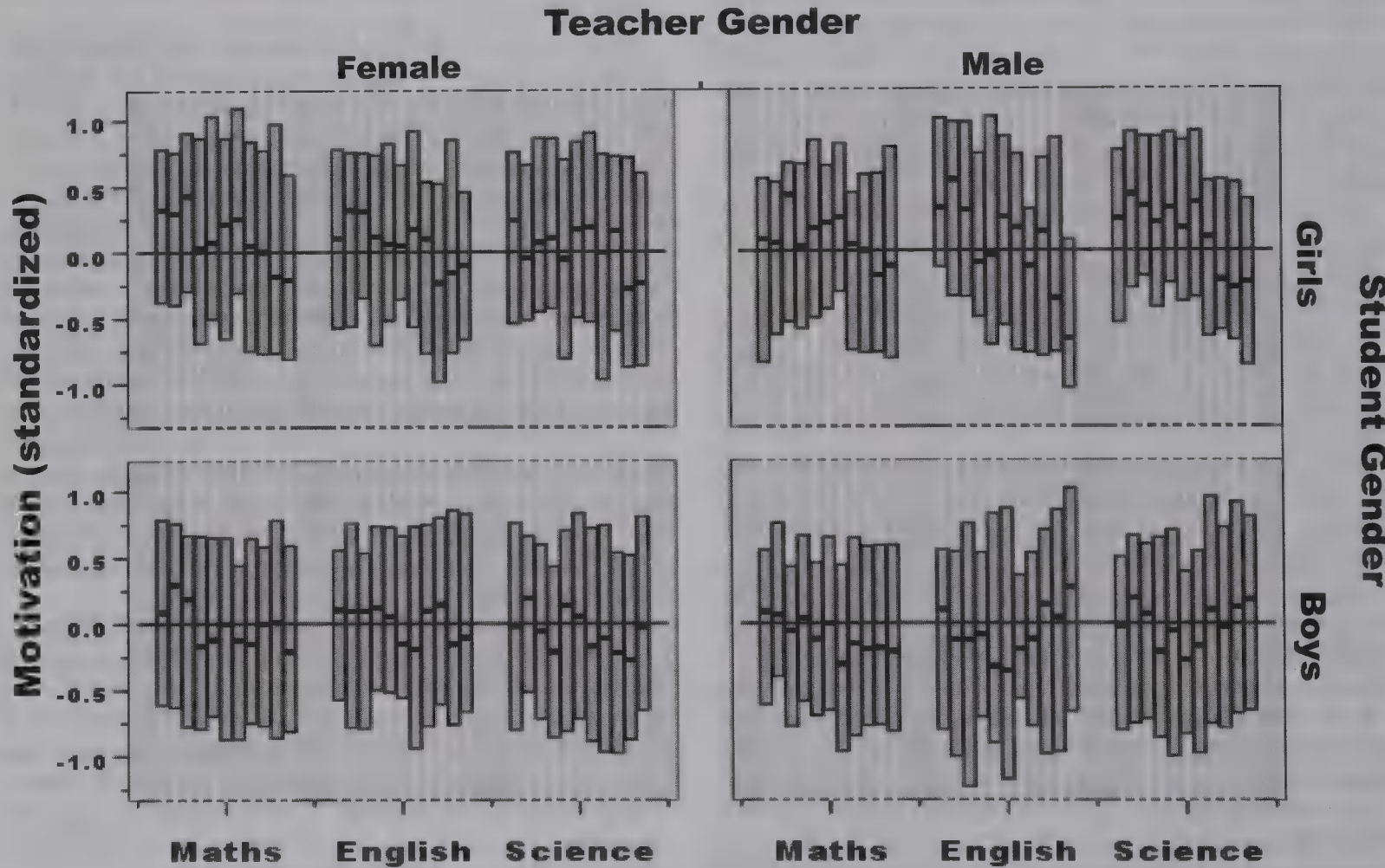


Figure 1. Box plots of 11 motivation scales for each of three school subjects, presented separately as a function of student and teacher gender. Box plots represent the median (dark line within each box) and the 25th and 75th percentiles (the boxes). The 11 motivation scales (in order of presentation) were Efficacy, Value, Mastery, Plan, Manage, Persist, Anxiety, Uncertain Control, Avoid, Handicapping, and Disengage.

teachers. Hence, these preliminary analyses do not seem to provide either support for gender-stereotypic differences in motivation in different school subjects or support for the matching hypothesis that boys are better motivated by male teachers, but these analyses seem to support the gender invariance and similarities model.

### *Cross-Classified Multilevel Analyses*

To more appropriately evaluate relations between independent variables (student gender, teacher gender, year in school, school subject, and all their interactions—represented by a set of 23 dummy variables) and dependent variables (adaptive motivations, maladaptive motivations, affective outcomes, and corresponding perceptions of classroom climate), we estimated a series of cross-classified models for each dependent variable. The effects of the entire set of predictor variables were estimated for individual students' ratings of their own motivation (Model 1; see Tables 1, 2, and 3) and their perceptions of the classroom climate. In preliminary analyses, variance component models (with no independent predictor variables, sometimes referred to as "null" models) were estimated separately for individual student ratings of their own motivation (the null model corresponding to Model 1 [M1o]) and their corresponding classroom climate perceptions (M2o). Finally, in Model 3, we added the class average of classroom climate ratings to Model 1 to determine how much of the variance in individual student motivation levels could be explained in terms of a generalized perception of classroom climate and how the inclusion of the class-average climate ratings influenced other fixed and random effects. Because of the complicated nature of these results, we present the results for self-efficacy (see Table 1) in detail and then summarize the key findings for other dependent variables.

**Self-efficacy.** Variance components for student self-efficacy (M1o; see the bottom of the first column in Table 1 under the heading *Efficacy*, labeled *Variance components*) show that variance components associated with the school, the teacher, and the classroom were all nonsignificant. This means the differences in efficacy did not differ significantly as a function of school, the teacher, or the classroom. The largest variance component (.573) was for the individual students. The means that a majority of the variance in self-efficacy ratings—approximately 57%—was due to differences between students. This is substantively important in that it shows that self-efficacy ratings by the same student generalized reasonably well over different (math, English, and science) subjects. As is typical in most psychosocial research, the residual variance component (which includes differences between subjects as well as a potentially infinite number of other sources of variance) was also statistically significant. Taken together, results of this variance components model (with no predictor variables) suggest that student self-efficacy was largely a function of the individual student and not the school they attended, their teacher, or the group of students in their class.

The pattern of results for variance components associated with self-efficacy classroom climate (M2o) was very different. Whereas variance components associated with school and teacher were not statistically significant, the variance component for classroom was significant. Although the variance component associated with individual students was still statistically significant, it was much smaller than was the case of students' ratings of their own self-

efficacy. For classroom climate perceptions, the largest variance component (.713) was the residual variance term that included differences associated with different subjects.

The juxtaposition between M1o and M2o supports the construct validity of the classroom climate perceptions. While self-efficacy of individual students should be largely a function of the individual student, classroom climate perceptions should be a function of the classroom and teacher rather than the individual student (see the earlier discussion). The results support these expectations. Individual student self-efficacies generalized over subjects, whereas classroom climate perceptions of self-efficacy did not. Although the variance component associated with the class was not large, it should be noted that the reliability of the aggregated class-average perception of climate was a function of the variance component and the number of students in the class. Hence, a modest variance component associated with class can result in a reasonably reliable measure of classroom climate when based on responses from a sufficient number of students. It is important to note, however, that the classroom climate did not appear to be a function of the teacher—only the group of other students in the classroom. Hence, the classroom climate associated with a particular class did not generalize to other classes taught by the same teacher.

In Model 1, individual student self-efficacies were related to a set of 23 dummy variables representing the 4 independent variables: student gender (boy, a z-scored dummy variable in which positive values reflected higher scores for boys), teacher gender (male), year in school (Year 10), subject (math and science, with English as the "left out" level in the traditional set of dummy variables) and all possible interactions (see Table 1). Only 2 of 23 effects were statistically significant, and each of these was small. Girls had higher self-efficacies (averaged across the three subjects), but the difference was not large (approximately .09 standard deviations, as all the variables were standardized to facilitate interpretations). There was also a three-way interaction between year, student gender, and teacher gender. Supplemental analyses showed that whereas both boys and girls had higher self-efficacies when taught by women than men in Year 10, in Year 8 girls had similar self-efficacies in classes taught by men and women but boys had substantially higher self-efficacies for classes taught by women than for those taught by men. Hence, the nature of this interaction does not support the benefit of male teachers for boys. There were no differences in self-efficacy for different school subjects, nor did school subject interact with any of the other variables. The residual variance components (random effects in Table 1) were very similar to those in the variance components model (M1o), indicating that very little of the variance at any level can be explained in terms of the 23 independent variables considered in Model 1 (Table 1).

In Model 2 (M2 in Table 1), a parallel analysis was conducted, with classroom climate perceptions of self-efficacy as the dependent variable. Only 1 of 23 effects was even marginally significant ( $.01 < p < .05$ )—the Science  $\times$  Student Gender  $\times$  Teacher Gender interaction. In contradiction to predictions based on the matching hypothesis, girls rated the self-efficacy class climate higher for science classes taught by men than for those taught by women, whereas boys gave higher ratings for classes taught by women than for those taught by men. Again, the residual variance components were nearly the same as the variance components in M2o, indicating that little variance was explained by the set of 23 predictor variables.



Table 1  
Adaptive Motivations: Effects of Student Gender, Teacher Gender, School Subject, Year in School, and Their Interactions on Individual Student Motivation (Model 1), Classroom Climate (Model 2), and Their Combination (Model 3)

Variable	Efficacy			Value			Mastery			Planning			Manage			Persist		
	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
Boy student (BS)	-.09**	.02	-.09**	-.07**	.01	-.07**	-.17**	-.01	-.17**	-.08**	.07*	-.08**	-.11**	.00	-.11**	-.11**	.04	-.10**
Male teacher (MT)	-.03	-.04	-.01	-.04	-.01	-.04	-.06	.01	-.05	.01	.01	.01	-.02	-.02	-.02	-.01	-.06	.01
BS × MT	.04	.05	.04	.04	.07**	.04	.02	.03	.01	.03	.01	.03	.02	.03	.02	.00	.05*	.01
Math (M)	.01	.03	.00	.01	.03	.01	.09*	-.07	.11**	-.04	-.03	-.03	.02	.00	.02	.00	.03	.01
M × BS	.06	-.04	-.06*	.07*	-.05	.07*	.03	-.06	.03	-.02	.02	-.02	-.05*	-.07*	-.05*	.01	-.04	.00
M × MT	-.04	-.05	.04	-.03	-.01	-.03	-.01	-.02	-.01	.03	.06	.01	.00	.06	.00	.00	-.02	.01
M × BS × MT	.03	-.05	.04	.02	-.05	.03	.02	-.06	.02	.01	.03	.01	.00	-.03	.00	-.01	.01	-.01
Science (S)	.04	.04	.03	-.04	-.03	-.01	.14**	-.03	.14**	-.01	-.03	-.01	.05	.00	.05	.01	.02	.01
S × BS	.03	-.06	.03	.01	-.08*	.01	.00	-.09**	.01	-.05	.02	-.06	-.04	-.08*	-.04	-.01	-.04	-.02
S × MT	-.03	.01	-.05	-.02	.03	-.03	.02	.00	.03	.01	-.01	.01	.01	.00	.01	-.04	-.03	-.04
S × BS × MT	.00	-.07*	.01	-.02	-.09*	-.02	.02	-.06	.02	.00	.00	.01	-.03	-.08	-.03	-.03	-.03	-.03
Year 10 (Y10)	.03	.02	.03	-.07*	.02	-.05	.05	-.02	.05	-.03	.01	-.03	.00	.00	-.01	-.03	.06	-.06
Y10 × BS	-.01	.01	-.01	.01	.02	.01	-.01	.02	-.01	.02	.01	.03	.02	.02	.02	.00	.00	.00
Y10 × MT	-.03	-.03	-.03	-.07*	.02	-.06*	-.07*	-.03	-.07*	.00	.00	.00	-.02	-.06	-.01	-.02	-.04	-.02
Y10 × BS × MT	-.07**	-.03	-.06**	-.07**	.03	-.07**	-.05*	-.01	-.05*	-.01	-.03	-.01	-.02	.02	-.02	-.03	-.02	-.03
Y10 × M	-.01	.10	-.05	.02	.03	.01	.07*	.04	.06	.04	.07	.02	.01	.07	.00	.01	.08	-.01
Y10 × M × BS	.00	.08	.00	.00	-.01	.00	.00	.03	.00	-.01	.02	-.02	.06*	-.01	.06*	-.04	-.01	-.04
Y10 × M × MT	.05	.03	.02	.07	.02	.05	.13**	.03	.12**	-.05	.04	-.06	.02	.00	.02	.00	.01	-.01
Y10 × M × BS × MT	-.01	.05	.01	-.01	-.04	-.01	.00	-.03	.01	-.02	.00	-.02	-.04	-.01	-.04	-.04	-.05	-.04
Y10 × S	-.06	.10	-.08*	-.01	.02	-.02	.02	.01	.02	.03	.02	.02	-.03	.03	-.03	-.05	.01	-.04
Y10 × S × BS	.02	.04	.02	.05	.00	.05	.03	.02	.03	.02	.01	.01	-.06	.00	-.06	-.03	.01	-.03
Y10 × S × MT	.07	.01	.02	.09*	.08	.06	.13*	.06	.12*	-.02	.07	-.04	.03	.06	.03	.03	.05	.01
Y10 × S × BS × MT	.01	-.06	.01	-.03	-.08*	-.03	-.01	-.05*	-.01	-.03	-.04	-.04	-.06*	-.02	-.06*	-.03	-.03	-.03
Climate			.36**			.39**			.21**			.31**			.13**			.37**

(Table continues)

Table 1 (continued)

Variable	Efficacy			Value			Mastery			Planning			Manage			Persist		
	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
Random effects																		
School	.031	.026	.022	.012	.015	.010	.025	.014	.017	.008	.014	.005	.013	.014	.011	.021	.017	.010
Teacher	.016	.020	.007	.037	.018	.012	.010	.013	.009	.008	.013	.008	.004	.013	.005	.008	.021	.006
Classroom	.012	.121**	.005	.013	.104**	.010	.024	.072*	.020	.013	.072**	.008	.005	.072**	.004	.018	.098**	.006*
Student	.582**	.147**	.552**	.527**	.220**	.514**	.615**	.253*	.608**	.627**	.253**	.617**	.711**	.253**	.704**	.596**	.240*	.568***
Residual	.369**	.708**	.369**	.404**	.667**	.413**	.304**	.673*	.308**	.355**	.673**	.359**	.278**	.673**	.281**	.366**	.633*	.379***
Deviance	3,129	4,236	3,189	3,284	4,135	3,119	2,797	4,150	2,817	3,062	4,150	3,078	2,747	4,150	2,665	3,115	4,045	3,171
Variance components	M1o	M2o		M1o	M2o		M1o	M2o		M1o	M2o		M1o	M2o		M1o	M2o	
School	.029	.012		.012	.008		.025	.010		.007	.010		.015	.010		.017	.009	
Teacher	.012	.026		.025	.022		.014	.013		.008	.013		.004	.021		.006	.028	
Classroom	.016	.115**		.026	.085**		.043**	.062**		.008	.062**		.004	.071**		.015	.090**	
Student	.573**	.145**		.513**	.209**		.623**	.244**		.644**	.244**		.718**	.204**		.610**	.239**	
Residual	.384**	.713**		.431**	.683**		.312**	.681**		.350**	.681*		.283**	.704**		.365**	.636**	
Deviance	3,197	4,249		3,391	4,176		2,840	4,169		3,037	4,169		2,674	4,227		3,107	4,053	

Note. Each dependent variable (the six adaptive motivations) was related to the set of 23 predictor variables (fixed effects; dummy variables representing student gender, teacher gender, year in school, school subject, and all possible interactions) and was included in a set of three models: Model 1 (M1; student self-ratings of their motivation; also see the corresponding variance components model M1o), Model 2 (M2; student self-ratings of classroom climate; also see the corresponding variance components model M2o), and Model 3 (M3; M1 with the addition of class-average ratings of the corresponding classroom climate). All analyses were based on a cross-classified multilevel model with five levels, the random effects: school, teacher, classroom, student, and subject (residual variance term).

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .



In Model 3, class-average values of classroom perceptions, which were the focus of Model 2, were added to Model 1. There was a large, statistically significant effect of classroom climate on self-efficacy perceptions; self-efficacies were higher in classes in which the classroom climate supported self-efficacy. However, other effects were largely similar to those in Model 1. It is important to note that the residual variance components were somewhat smaller in Model 3 (compared to Model 1 and M1o) because classroom climate had a substantial effect on individual student self-efficacies.

In summary, the self-efficacy results provide no support for predictions based on matching, gender-stereotypic, and gender intensification perspectives (i.e., predictions that boys are better motivated by male teachers). Girls had higher self-efficacies in all subjects, and this gender difference did not vary with subject or age. Although a few Student Gender  $\times$  Teacher Gender interactions were significant, the nature of these interactions was in the opposite direction from that predicted by the matching hypothesis—particularly the predicted benefits of male teachers for boys. Broadly, these results support the gender invariance and similarities model.

*Other adaptive motivations.* The same set of models was applied to the remaining five adaptive motivations (see Table 1). Again, residual variance components in Models 1 and 2 were similar to variance components in M1o and M2o, indicating that the set of 23 predictors did not explain much variance in any of the motivation and classroom climate perceptions.

In Models 1 and 3, girls consistently had higher levels of motivation than boys, but the Teacher Gender  $\times$  Student Gender interactions were nonsignificant for all six adaptive motivations. However, the Year  $\times$  Student Gender  $\times$  Teacher Gender interaction was significant for valuing and marginal ( $.01 < p < .05$ ) for mastery. Girls had similar mastery motivations in classes taught by male and female teachers in both Years 8 and 10, whereas boys had higher mastery motivation in classes taught by female teachers than in classes taught by male teachers in Year 10 but somewhat higher motivation in classes taught by male teachers than in classes taught by female teachers in Year 8. Mastery motivation was higher in science and math than in English but was particularly low in English classes for older students taught by men. However, this effect was similar for boys and girls.

For classroom climate ratings (Model 2), there were few statistically significant effects for any of the other adaptive motivations; only 7 of 115 (23 effects  $\times$  5 motivations) were statistically significant at the nominal .05 level. However, the Student Gender  $\times$  Teacher Gender interaction was significant for two of these scales (value and persistence). For persistence class climate, the effect was marginal ( $.05 < p < .01$ ); ratings of classes taught by women were higher than ratings of classes taught by men for both boys and girls. For male teachers, however, ratings were higher for boys than for girls. For value classroom climate, the Student Gender  $\times$  Teacher Gender interaction varied by subject and year and was only significant for English classes. Whereas climate in English classes taught by female teachers was perceived similarly by boys and girls in both year groups, classes taught by male teachers were rated higher by boys in Year 8 than by boys in Year 10 and lower by girls in Year 10 than by girls in Year 8.

*Maladaptive motivations.* The same set of models was applied to the set of five maladaptive motivations (see Table 2). Variance components for the school and teacher were nonsignificant,

whereas variance associated with the class was significant for four of five classroom climate ratings (all but avoidance) and for two of the motivation scales (handicapping and disengagement). With the possible exception of anxiety, the residual variance components in Model 1 and Model 2 were similar to variance components in M1o and M2o, indicating that the set of predictors did not explain much variance in any of these maladaptive motivation and classroom climate perceptions.

Fixed effects were typically nonsignificant and small for the set of maladaptive scales other than anxiety. For individual student motivation (Models 1 and 3), anxiety was greater in Year 10 than in Year 8. There was a moderate effect of student gender in that girls reported being more anxious than boys, but this student gender difference did not depend on the gender of the teacher. Whereas anxiety was marginally higher in math, this difference was not large. Classroom climates were more anxious in math classes (particularly for girls), science classes, and Year 10. Across all five maladaptive motivations, there were no statistically significant interactions between student gender and teacher gender, and this lack of interaction generalized over school subject and year level.

*Affective outcomes.* The same set of models was applied to the set of five affective outcomes (see Table 3). Unlike earlier analyses, there were statistically significant variance components associated with the teacher for both individual student and climate perceptions of enjoyment and particularly student-teacher relationships, although class-level variance components were also significant. In addition, there were significant class-level effects for two additional individual scales (future and relationship) as well as for all five climate scales.

Given the diverse nature of these affective outcomes, it is not surprising that the pattern of fixed effects did not generalize well across the different outcomes. Important for purposes of the present investigation, there was only one Student Gender  $\times$  Teacher Gender interaction for individual student outcomes (for student-teacher relationships), and the direction of this interaction was not entirely consistent with a matching hypothesis—particularly the benefits of male teachers for boys. Both boys and girls rated their student-teacher relationships as being better in classes taught by female teachers than in classes taught by male teachers. However, the difference in favor of female teachers was much larger for girls than for boys. The nature of this interaction did not vary significantly with school subject and year in school.

Classroom climate perceptions of participation and student-teacher relationships varied in complicated functions of teacher gender, year, and subject. For participation, this was due primarily to the very low climate perceptions in Year 10 English classes taught by men. For student-teacher relationships in Year 10 science, climate perceptions were very high for male teachers, whereas in Year 10 English, perceptions were higher for female teachers than for male teachers. However, none of these effects varied with student gender. More generally, outcomes tended to be more negative in mathematics (enjoyment, future, and resilience) and science (relationship and resilience). While these differences were evident to varying degrees in both individual and class-climate scales, they varied to some extent according to year in school.

In Model 3, there were substantial positive effects of class-average climate scores for each of the individual scales. Indeed, after class-average perceptions of climate were controlled for each

Table 2

*Maladaptive Motivations: Effects of Student Gender, Teacher Gender, School Subject, Year in School, and Their Interactions on Individual Student Motivation (Model 1), Classroom Climate (Model 2), and Their Combination (Model 3)*

Variable	Anxiety			Unc control			Avoid			Handicapping			Disengage		
	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
Fixed effects															
Boy student (BS)	-.24**	-.03	-.24**	-.03	-.01	-.03	-.02	.03	-.02	.02	-.01	.02	.04	.02	.04
Male teacher (MT)	-.02	-.08*	-.02	-.01	.04	-.02	.02	.00	.02	-.01	.01	-.02	.02	.01	.02
BS × MT	.00	.01	.00	.04	-.04	.04	.01	.00	.01	.02	-.02	.02	-.01	.00	-.01
Math (M)	.07*	.20**	.06	.06	-.04	.07	-.03	.10*	-.04	.04	-.02	.04	-.05	.00	-.06
M × BS	-.04	-.09**	-.04	-.04	-.03	-.04	-.05	-.02	-.05*	-.01	.05	-.01	-.07*	-.07*	-.07*
M × MT	-.01	.02	-.01	.00	.01	.00	-.04	-.02	-.04	-.01	.01	-.01	.01	-.02	.01
M × BS × MT	-.03	-.02	-.03	.00	.03	.00	-.02	.02	-.02	-.02	.04	-.01	-.04	-.01	-.04
Science (S)	.08	.13*	.08*	.07	-.06	.08	-.02	.02	-.03	-.02	-.03	-.01	-.04	-.07	-.02
S × BS	-.01	-.04	-.01	-.01	.00	-.01	-.04	-.01	-.04	.01	.05	.01	.00	-.03	.00
S × MT	.05	.03	.05	.05	.03	.05	-.01	-.05	-.01	.05	.05	.04	.02	-.01	.03
S × BS × MT	-.03	-.07	-.03	.00	.04	.01	-.01	.00	-.01	-.01	.06	-.01	-.01	.02	-.01
Year 10 (Y10)	.11**	.10**	.10**	.02	.00	.02	-.02	.03	-.03	-.04	.00	-.04	-.07*	.06	-.05
Y10 × BS	-.04	-.01	-.04	-.03	-.02	-.03	-.05	-.05	-.05	.00	-.03	.00	-.01	-.02	-.01
Y10 × MT	.01	-.05	.02	.04	.02	.03	-.02	.01	-.01	.06	.02	.06*	.04	.06	.02
Y10 × BS × MT	.00	-.01	.00	-.02	.05	-.02	.02	-.02	.02	.02	-.05	.01	.03	.00	.02
Y10 × M	-.04	.03	-.04	-.07	-.04	-.06	-.02	.00	-.02	-.07	.04	-.09*	-.06	.00	-.06
Y10 × M × BS	-.04	-.04	-.04	.00	-.02	.00	.00	-.04	.00	.00	-.07	.00	.08*	-.03	.08**
Y10 × M × MT	-.04	-.08	-.03	-.05	-.07	-.03	-.02	-.05	-.01	-.03	-.07	-.01	-.09	-.09	-.04
Y10 × M × BS × MT	-.03	-.07	-.03	-.01	.00	-.01	-.01	-.02	-.01	.01	.01	.01	.02	.00	.02
Y10 × S	-.07*	.04	-.07*	-.08*	.00	-.08*	-.05	-.03	-.04	-.01	.09	-.05	.01	.00	-.01
Y10 × S × BS	-.03	-.04	-.03	.00	-.01	.00	.00	-.04	.00	-.05	-.04	-.05	.05	-.01	.05
Y10 × S × MT	.01	.00	-.01	.03	-.05	-.01	.01	-.01	-.01	-.06	-.01	-.06	-.10*	-.07	-.06*
Y10 × S × BS × MT	.00	-.02	.00	.03	-.02	.03	.01	-.08*	.01	.02	-.03	.02	.04	.01	.04
Climate			.06			.29**			.16*			.38**			.48**
Random effects															
School	.015	.011	.013	.062	.023	.031	.016	.015	.013	.066	.018	.039	.027	.017	.012
Teacher	.003	.014	.003	.018	.011	.010	.006	.021	.006	.010	.012	.009	.018	.015	.007
Classroom	.003	.048*	.003	.016	.043*	.016	.006	.022	.005	.047**	.075**	.033**	.031	.100**	.007
Student	.672**	.232**	.671**	.623**	.146**	.613**	.678**	.319**	.673**	.615**	.213**	.605**	.524**	.145**	.524**
Residual	.259**	.672**	.260**	.339**	.808**	.345**	.313**	.743**	.315**	.311**	.701**	.316**	.407**	.745**	.407**
Deviance	2,528	4,146	2,534	2,981	4,462	2,301	2,846	4,299	2,860	2,838	4,219	2,864	3,295	4,307	3,295
Variance components	M1o	M2o		M1o	M2o		M1o	M2o		M1o	M2o		M1o	M2o	
School	.020	.012		.052	.022		.012	.013		.069	.014		.026	.020	
Teacher	.007	.029		.019	.010		.007	.025		.012	.011		.015	.013	
Classroom	.005	.074**		.013	.036*		.004	.020		.044**	.069**		.033*	.091**	
Student	.731**	.232**		.620**	.140**		.674**	.213		.615**	.210**		.523**	.153**	
Residual	.265**	.672**		.342**	.813**		.314**	.745**		.312**	.708**		.418**	.738**	
Deviance	2,566	4,147		3,000	4,472		2,851	4,324		2,843	4,237		3,340	4,307	

*Note.* Each dependent variable (the five maladaptive motivations) was related to the set of 23 predictor variables (fixed effects; dummy variables representing student gender, teacher gender, year in school, school subject, and all possible interactions) and was included in a set of three models: Model 1 (M1; student self-ratings of their motivation; also see the corresponding variance components model M1o), Model 2 (M2; student self-ratings of classroom climate; also see the corresponding variance components model M2o), and Model 3 (M3; M1 with the addition of class-average ratings of the corresponding classroom climate). All analyses were based on a cross-classified multilevel model with five levels, the random effects: school, teacher, classroom, student, and subject (residual variance term). Unc = uncertain.

\*  $p < .05$ . \*\*  $p < .01$ .

of these affective outcomes, the statistically significant variance components associated with teacher and/or class levels (in Model 1) were no longer significant. There was, however, substantial variation in the sizes of effects for class-average perceptions across the different outcomes, varying from a very large effect of .85 for student-teacher relationships to .69 and .62 for future and enjoyment and .43 and .30 for participate and resilience. In general,

controlling for class-average climate perceptions eliminated the statistically significant variance components associated with teacher and class levels and most of the fixed effects observed in Model 1. Hence, many of the generally small effects on student motivation associated with student gender, teacher gender, school subject, and year in school were mediated by the corresponding measure of classroom climate.



Table 3  
Educational Outcomes: Effects of Student Gender, Teacher Gender, School Subject, Year in School, and Their Interactions on Individual Student Motivation (Model 1), Classroom Climate (Model 2), and Their Combination (Model 3)

Variable	Participate			Enjoy			Future			Relationship			Resilience		
	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
Fixed effects															
Boy student (BS)	-.07*	-.05	-.06*	-.07**	.01	-.07**	-.03	-.01	-.02	-.06**	-.03	-.06*	.14**	-.01	.14**
Male teacher (MT)	-.01	-.03	.01	.01	.04	-.01	.01	-.02	.02	-.05	.00	-.04	.00	-.02	.01
BS × MT	.03	-.02	.03	.05	.06**	.04	.04	.03	.04	.08**	-.06*	.08**	.02	.03	.02
Math (M)	.00	.00	.00	-.12*	-.24**	-.05	-.17*	-.12*	-.07*	-.13	-.07	-.05	-.08*	-.12*	-.06*
M × BS	.01	.03	.01	.01	-.02	-.01	.03	.01	.03	.01	.00	.00	.03	.01	.03
M × MT	.02	.04	-.01	.03	-.04	-.01	-.09	.01	-.09*	-.07	-.05	-.04	-.08*	.01	-.08*
M × BS × MT	.01	-.02	.01	.02	-.05	.02	.00	-.03	.00	.00	-.05	.00	.05	-.03	.05
Science (S)	.04	.06	.01	-.07	-.07	-.02	-.10	-.03	-.06	-.19*	-.13	-.05	-.08*	-.03	-.07
S × BS	-.02	.02	-.01	-.03	-.04	-.03	.00	-.04	.00	-.02	-.02	-.03	-.01	-.04	-.01
S × MT	.05	.06	.01	-.01	.01	-.01	-.06	.01	-.08	-.05	.01	-.06	-.08*	.01	-.08*
S × BS × MT	-.06	-.01	-.05	-.01	-.06	-.01	-.02	-.03	-.02	-.03	-.03	-.03	.03	-.03	.03
Year 10 (Y10)	.04	.02	.04	.02	.03	.01	.00	.09*	-.06*	.04	.07	.02	-.10**	.09*	-.10**
Y10 × BS	.02	.06	.02	.02	.01	.02	.02	-.01	.03	.02	-.01	.02	.04	-.01	.05
Y10 × MT	-.01	-.04	.01	-.04	-.03	-.03	-.05	-.04	.00	-.02	.01	-.02	.01	-.04	.01
Y10 × BS × MT	-.03	.00	-.03	-.05	-.04	-.05	-.05	-.02	-.04	-.03	-.01	-.05	-.02	-.02	-.02
Y10 × M	.02	.13**	-.03	.06	.06	.03	.05	.13*	-.04	.07	.06	.02	-.02	.13*	-.03
Y10 × M × BS	.02	.04	.02	-.01	.02	-.01	-.01	.05	-.02	-.01	-.03	.00	.01	.05	.01
Y10 × M × MT	.09*	.13*	.02	.08	.08	-.01	.06	.04	.03	.15*	.10*	.06	-.01	.04	.00
Y10 × M × BS × MT	-.05	.00	-.05	-.02	-.03	.00	-.03	-.02	-.03	-.06	-.05	-.04	-.01	-.02	-.01
Y10 × S	-.05	.09	.00	.01	.04	.00	-.06	.06	-.10**	.09	.10	.01	.01	.06	.00
Y10 × S × BS	-.01	.02	.00	.00	.04	.00	.01	.07*	.00	-.02	.00	-.01	.02	.07*	.01
Y10 × S × MT	.09*	.18**	.01	.07	.09	.01	.06	.09	.02	.18**	.13*	.07	-.02	.09	-.02
Y10 × S × BS × MT	-.05	.01	-.05	-.02	-.04	-.02	-.04	-.04	-.03	-.04	-.04	-.03	-.03	-.04	-.03
Climate			.43**			.62**			.69**			.85**			.30**
Random effects															
School	.028	.013	.015	.016	.015	.007	.019	.016	.009	.032	.034	.004	.026	.010	.019
Teacher	.020	.023	.006	.057*	.103*	.002	.053	.026	.008	.114*	.155**	.006*	.006	.008	.004
Classroom	.015	.073**	.006	.017	.083*	.003	.030	.097**	.005	.115*	.106*	.005*	.007	.049**	.005
Student	.615**	.241**	.601**	.443**	.214**	.414**	.349**	.194**	.323**	.221**	.082*	.212**	.686**	.218**	.672**
Residual	.344**	.653**	.351**	.474**	.582**	.478**	.558**	.672**	.569**	.545**	.667*	.548**	.293**	.743**	.298**
Deviance	3,010	4,099	3,043	3,553	3,902	3,567	3,831	4,148	3,865	3,793	4,139	3,802	2,734	4,318	2,761
Variance components	M1o	M2o		M1o	M2o		M1o	M2o		M1o	M2o		M1o	M2o	
School	.015	.010		.009	.014		.011	.011		.013	.014		.030	.006	
Teacher	.017	.020		.058*	.132**		.046	.025		.115*	.127*		.008	.007	
Classroom	.016	.083**		.020	.081*		.054*	.121**		.132**	.131**		.008	.040*	
Student	.618**	.244**		.437**	.202**		.334**	.189**		.211**	.081*		.713**	.215**	
Residual	.350**	.651**		.482**	.593**		.571**	.675**		.562**	.672**		.293**	.743**	
Deviance	3,039	4,094		3,582	3,935		3,869	4,155		3,843	4,149		2,732	4,317	

Note. Each dependent variable (the five outcomes) was related to the set of 23 predictor variables (fixed effects; dummy variables representing student gender, teacher gender, year in school, school subject, and all possible interactions) and was included in a set of three models: Model 1 (M1; student self-ratings of their motivation; also see the corresponding variance components model M1o), Model 2 (M2; student self-ratings of classroom climate; also see the corresponding variance components model M2o), and Model 3 (M3; M1 with the addition of class-average ratings of the corresponding classroom climate). All analyses were based on a cross-classified multilevel model with five levels, the random effects: school, teacher, classroom, student, and subject (residual variance term).

\*  $p < .05$ . \*\*  $p < .01$ .

Discussion

Gender-Stereotypic Differences, Gender Intensification, and the Matching Hypothesis

Substantively, the overarching focus of the study is whether male teachers are better able to motivate boys, how these effects

vary with year in school and school subject, and the role of classroom climate. We began by reviewing various predictions based on the gender-stereotypic model, the gender intensification model, and the matching hypothesis. It was predicted that boys and girls would be more motivated in gender-stereotypic disciplines (girls in English, boys in math and science) and that these differ-

ences would grow larger with age. According to the matching hypothesis as applied in our study, male teachers should be better able to motivate boys in particular, whereas female teachers should also be better able to motivate girls. However, across a set of cross-classification MLMs applied to six adaptive motivations, five maladaptive motivations, and five affective outcomes, there was little or no support for these predictions. Instead, there was better support overall for the gender invariant and similarities model. There were reasonably consistent gender differences in favor of girls. More important and consistent with Marsh's (1993) invariance hypothesis, the effects of student gender did not vary substantially as a function of school subject, age, or teacher gender.

Consistent with expectations, for all six adaptive motivations, girls were generally more positively motivated than boys. However, contrary to predictions based on the gender-stereotypic model (but consistent with the gender invariance hypothesis), these gender differences in favor of girls generalized across school subjects. Indeed, gender differences favoring girls were as large in math and science as in English (or even larger). Furthermore, these gender differences in favor of girls in different school subjects generalized over age. For the five maladaptive motivations, there were only substantial gender differences in one scale. Girls were considerably more anxious than boys, but these differences did not vary as a function of year in school or school subject. For the remaining maladaptive motivations, there were no significant differences between boys and girls, and this lack of gender difference generalized reasonably well over year in school and school subject. For the set of six outcome variables, girls had more positive outcomes for participation, enjoyment, and student-teacher relationships, but boys had higher resilience scores.

Particularly central to the present investigation were the Student Gender  $\times$  Teacher Gender interactions and predictions that boys would be benefited by having male teachers. There was, however, no support at all for the matching hypothesis for any of the adaptive or maladaptive motivations. Although there were small gender differences (mostly in favor of girls) in some of the motivations, these did not depend on the gender of the teacher. The only statistically significant interaction between student gender and teacher gender for individual student constructs was for student-teacher relationships. For this one outcome, there was partial support for a matching hypothesis. Even here, however, both boys and, particularly, girls felt they had better student-teacher relationships with female teachers than with male teachers. Hence, this does not support the contention that boys are benefited by male teachers; boys were more benefited by female teachers, but less so than were girls.

### *Classroom Climate*

We predicted that there would be statistically significant and substantively meaningful variance components associated with the teacher who taught a class and the particular group of students in a given class. Although these two sources of classroom climate are typically confounded, the design of our study (with multiple classes for each teacher and the same students making ratings of multiple teachers) and the MLM analyses allowed us to disentangle these two sources of variation in class-climate perceptions. The results showed that for both adaptive and maladaptive climate

perceptions there were consistent differences between classes but that these differences did not generalize over different classes taught by the same teacher. This finding has potentially profound implications for classroom climate research, suggesting that the classroom climate in relation to a wide variety of motivational constructs is largely a function of the particular group of students in a class rather than the teacher who teaches the class. We do note, however, that teacher-level variance components (as well as class-level components) were statistically significant for two of the affective outcome variables (enjoyment and student-teacher relationships). Hence, there were some aspects of classroom climate for which the particular teacher made a difference beyond the effects of the particular group of students in the class being taught. We also note that earlier research (Papaioannou et al., 2004) in a physical educational setting showed that classroom motivation climate was more a function of the teacher than of the group of students. Hence, there is a need for further research to test the replicability and generalizability of our results.

Gender differences in classroom-climate perceptions tended to be smaller than the gender differences in the corresponding motivations and outcome variables. Indeed, the only main effect of gender across the 11 motivation and 5 outcome climates was a small effect in favor of boys for planning (even though planning motivation was higher for girls). Even for anxiety and resiliency, for which girls had substantially less favorable scores than boys, there were no gender differences in the corresponding classroom-climate perceptions. Consistent with this finding, variance components associated with students were substantially smaller for class-climate perceptions than for corresponding student motivations. There were, however, a number of statistically significant effects of school subject and year in school on classroom-climate perceptions, which sometimes interacted with student gender. Girls perceived the management climate to be stronger in math and science classes than did boys. Anxiety classroom-climate perceptions were higher in mathematics and science than English and higher in Year 10 than Year 8, but girls perceived the anxiety and disengagement climates to be larger (more maladaptive) in math classes than did boys. In general, however, there was good agreement between boys and girls in terms of their perceptions of classroom climates.

### *Significance and Implications of Findings*

Counter to popular arguments that boys fare better academically under male teachers, there was little or no support for the matching hypothesis for either motivation or the outcome variables considered here. Taken as a whole, this argument supports previous contentions by MLM research (Hill & Rowe, 1996; MacDonald, Saunders, & Benfield, 1999; Rowe, 2000; Rowe & Rowe, 2002) that where there is a class or teacher impact on academic outcomes, it is probably the nature of pedagogy that is critical and not the gender of the person delivering it. This is also consistent with previous qualitative work (Martin, 2002, 2003a; Martino & Meyenn, 2002) that has shown that it is the nature of pedagogy, rather than demographic-type variables, that students are most concerned about.

Gender differences were not large and, where significant, tended to favor girls. Anxiety and persistence, however, were exceptions to this generalization in that boys had more favorable scores on



these scales than girls. Although they are consistent with previous research (Martin, 2001, 2002, 2003a, 2003b, 2003c, 2003d, 2004, 2007; Martin & Marsh, 2005), our results based on a cross-classified MLM extend previous research. The education of boys has been an issue of ongoing debate, research, and policy implementation over the past decade (Weaver-Hightower, 2003). It appears that boys' and girls' educational outcomes are most divergent during the period of adolescence, with a large body of data attesting to the differences between adolescent boys and girls in academic motivation and achievement and the comparative maturity and conscientiousness of girls' approach to their studies during this time (e.g., Collins, Kenway, & McLeod, 2000; Lingard, Martino, Mills, & Bahr, 2002; Martin, 2004). In reviews attempting to identify strategies to enhance boys' flagging motivation, literacy, and achievement (Lingard et al., 2002; Martin, 2004), there is consistent reference to the importance of supportive and affirming relationships in promoting the educational outcomes of boys. Given this, it is interesting to note that in relation to the present data, it was on the teacher-student relationship measure that most variance existed at the teacher and class levels. Another key finding is that the bulk of variance in motivation occurred at the student level, indicating that the motivation level of an individual student generalized reasonably well over different classes and different school subjects. In contrast, as expected, student perceptions of classroom climate were much more specific to a particular class taught by a particular teacher. Whereas there was weak evidence in support of class and teacher effects for individual student motivation, these effects were stronger for classroom-climate perceptions. However, even for motivational climate perceptions where class-level effects were significant, there was little evidence that the effects of a particular teacher in one class generalized to other classes taught by the same teacher. It was only with two affective outcomes (enjoyment and student-teacher relationships) that there were significant teacher effects—in addition to class effects. Findings, then, hold implications for educational intervention, suggesting that researchers should focus on individual students in addition to the teacher and the class composition.

Notwithstanding the substantial variance at the student level on individual measures, it is clear that teachers had a significant effect at the climate level—specifically enjoyment and student-teacher relationship. Indeed, a variety of studies reflect this. Students who believe their teacher is a caring one also tend to believe they learn more (Teven & McCroskey, 1997). Students' feelings of acceptance by teachers are associated with emotional, cognitive, and behavioral engagement in class (Connell & Wellborn, 1991). Specific strategies teachers can use to develop motivation and engagement in the classroom were detailed in Martin (2001, 2005, 2006b, in press; see also Martin & Marsh, 2005). Also raised by Martin (2006b) in previous teacher-related motivation and engagement research was the role of professional development in enhancing students' motivation and engagement.

Findings also suggest that the composition of the class is relevant to students' motivation and engagement. Indeed, given that in recent years there has been substantial focus on teacher effectiveness and the development of taxonomies and criteria of effective teaching, it might now be important to revisit the class composition. In particular, there is now a need for research focusing on the characteristics of effective classrooms (including but not restricted to characteristics of effective teaching), the students collected

together in the classroom, and the bases on which they are collected together. Obviously, prior work has been conducted into cognate issues, such as seating arrangement (Hastings & Schwieso, 1995; Marx, Fuhrer, & Hartig, 1999), streaming (Marsh, 1987a; Marsh & Hau, 2003), and single-sex class composition (Martin, 2004; Martin & Marsh, 2005); however, class composition as it is relevant to motivation and engagement is an avenue for further research, and multilevel techniques are well placed to provide valuable input.

### *Limitations and Future Directions*

The present study provides much information on the role of student-, class-, and school-level variance in explaining motivation in mathematics, English, and science and also sheds further light on the role of student gender and teacher gender in contributing to students' motivation and perceptions of classroom climate. There are, however, a number of potential limitations that are important to consider when interpreting of findings that also provide direction for further research.

The data presented in this study were all self-reported, some based on self-report responses by individual students and others based on the class-average responses drawn from groups of students. Although this is a logical and defensible methodology in its own right, given the substantive focus, it is important to conduct research that examines the same constructs using data derived from additional sources—for example, from teachers and parents—and also using different methodologies and paradigms, such as structured interviews or observation (see Dowson & McInerney, 2003; Martin, Marsh, Williamson, & Debus, 2003) as well as other multimethod approaches to the issue (Marsh, Martin, & Hau, 2006). Inclusion of achievement data is also relevant.

It is important to recognize that our study was based on only five schools; this limits generalizability to other schools and is also weak in terms of testing variance components at the school level. Future research needs to be conducted that includes a larger number and a wider variety of schools. Furthermore, although we had a large number of teachers in the study, we did not have a large number of classes taught by the same teacher. For this reason, our conclusion that motivation and classroom climate are more a function of the class of students than the teacher should be interpreted cautiously. Indeed, because there were only small effects of either teacher or class—particularly in relation to motivation—it might be a moot issue as to how much of this effect is associated with the teacher rather than the particular group of students within a given class. Nevertheless, the approach used here is a potentially important contribution to further research into this issue—in relation to class climate, which was our focus, but also in relation to other outcomes, such as achievement test scores. Hence, it would be useful in future research to collect data for a larger number of different classes taught by the same teacher to provide a stronger test of our conclusion that classroom climate is as much as or more a function of the group of students rather than the teacher.

Further on this issue of data collection, it is important to recognize that the data were collected at one time point, so future longitudinal work is needed to explore the stability of constructs over time and to provide greater scope to partial out residual variance, which in this study was manifested in student-level variance. Employment of longitudinal research also has the poten-



tial to clarify and uncover the possible motivational fluctuations across time (Bong, 1996; Jacobs, Lanza, Osgood, Eccles, & Wigfield, 2002). It would be particularly useful to evaluate changes in motivational and classroom climate perceptions at critical educational-stage transitions (e.g., from junior high school to middle high school to senior high school), although the timing and importance of such transitions will vary with the particular school system.

A major finding in the present study was that there was little or no evidence to support the benefit of male teachers for boys. This finding runs counter to the gender-stereotypic model and some popular claims and beliefs that boys' academic development is dependent on there being ample presence of male teachers in their academic life. In relation to this, it is important to consider carefully the generalizability of this finding. First, it does not necessarily apply to younger children, because the data collected in this study were from middle and high school students only. Similar research is needed at the elementary school level to ascertain the generalizability of the finding to younger children. Second, the finding does not necessarily apply to other emotional and psychosocial dimensions of students' lives not considered here. It may well be, for example, that boys prefer male teachers when dealing with emotional and personal issues, just as girls may prefer female teachers (Martin, 2002, 2003a). Indeed, this might suggest that the debate about the need for male teachers in boys' lives (and female teachers in girls' lives) confuses teachers with role models. This points to the need for research that seeks to understand in greater detail the congruence (or lack thereof) between teachers and role models, how the two differ, how they are similar, the extent to which one can be the other, and the gender-relevant issues involved.

Our findings suggest that climate effects associated with a class are due to the particular group of students and their characteristics, as much as or more than the effects associated with the teachers who teach the class, which generalize over different classes taught by the same teacher. Although the findings are substantively important, with significant implications for class-climate research, it is important not to overgeneralize interpretations of these results. Because the innovative design and methodological approach used here are not common in classroom-climate research (or teacher effectiveness research more generally), it is important to test the generalizability of these findings with other constructs where teacher effects might generalize better over different classes (e.g., teacher effectiveness, achievement, and other outcomes). Indeed, our research showed that there were generalizable teacher effects for two of our outcome variables (enjoyment of the class and student-teacher relationships). Furthermore, research on students' evaluations of teacher effectiveness (Marsh, 1987b, 2007b) shows that ratings of overall teacher effectiveness and even the profile of responses to specific components (e.g., learning, enthusiasm, organization, rapport, workload difficulty) generalize very well over different classes taught by the same teacher. This suggests that teacher effects are likely to be larger and more generalizable for constructs more closely aligned to actual teacher behaviors than for constructs more closely aligned to individual student characteristics, such as those considered here.

## Conclusion

The present study used cross-classified MLMs to determine the contribution of student gender and teacher gender across junior and middle high school classes in math, English, and science and the relative salience of student, class, and school variance in boys' and girls' academic motivation and corresponding classroom climates. As expected, of the few significant main effects that emerged, most of them were in girls' favor. Motivation did not vary substantially for boys and girls as a function of the teacher's gender, thus supporting the gender invariant and similarities model and calling into question the gender-stereotypic model and matching hypothesis for student motivation. Findings also demonstrated that the bulk of variance in motivation occurred at the student level. Individual student measures of motivation yielded few statistically significant class-level variance components, and no teacher-level components were significant. Taken together, the findings of the present investigation hold substantive and methodological implications for researchers studying issues relevant to motivation and classroom climate and are also relevant to educators seeking to enhance educational outcomes that rely in large part on the extent to which their students are affectively, cognitively, and behaviorally engaged.

## References

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, 84, 261-271.
- Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology*, 80, 260-267.
- Australian Labor Party. (2004). *Making every day Father's Day: More male role models for boys* (policy document). Canberra, Australia: Author.
- Bong, M. (1996). Problems in academic motivation research and advantages and disadvantages of their solutions. *Contemporary Educational Psychology*, 21, 149-165.
- Browne, W. J. (2005). *MCMC estimation in MLwiN (Version 2)*. Bristol, England: Centre for Multilevel Modelling, University of Bristol.
- Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *Journal of Higher Education*, 71, 17-33.
- Collins, C., Kenway, J., & McLeod, J. (2000). *Factors influencing the educational performance of males and females in school and their initial destinations after leaving school*. Canberra, Australia: Department of Education, Training & Youth Affairs.
- Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy, and relatedness: A motivational analysis of self-system processes. In M. R. Gunnar & L. A. Sroufe (Eds.), *Self processes in development: Minnesota Symposium on Child Psychology* (Vol. 29, pp. 244-254). Hillsdale, NJ: Erlbaum.
- Dee, T. S. (2006). The why chromosome: How a teacher's gender affects boys and girls. *Next Education*, 6(4). Retrieved September 21, 2007, from <http://www.hoover.org/publications/ednext/3853842.html>
- Dowson, M., & McInerney, D. M. (2003). What do students say about their motivational goals? Towards a more complex and dynamic perspective on student motivation. *Contemporary Educational Psychology*, 28, 91-113.
- Eccles, J. S. (1987). Gender roles and achievement patterns: An expectancy value perspective. In J. M. Reinisch, L. A. Rosenblum, & S. A. Sanders (Eds.), *Masculinity/femininity: Basic perspectives* (pp. 240-280). New York: Oxford University Press.



- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53, 109–132.
- Feldman, K. A. (1992). College students' views of male and female college teachers. Part I-Evidence from social laboratory and experiments. *Research in Higher Education*, 33, 317–351.
- Feldman, K. A. (1993). College students' views of male and female college teachers: 2. Evidence from students' evaluations of their classroom teachers. *Research in Higher Education*, 34, 151–211.
- Feller, B. (2006, August 27). Study: Teacher's gender affects learning. *Washington Post*. Retrieved September 21, 2007, from <http://www.washingtonpost.com/wp-dyn/content/article/2006/08/27/AR2006082700273.html>
- Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Hodder Arnold.
- Green, J., Martin, A. J., & Marsh, H. W. (2006). *Motivation and engagement in English, mathematics and science high school subjects: A domain-specific construct validity approach*. Sydney, Australia: SELF Research Centre, University of Western Sydney.
- Hastings, N., & Schwieso, J. (1995). Tasks and tables: The effects of seating arrangements on task engagement in primary classrooms. *Educational Research*, 37, 279–291.
- Hill, J. P., & Lynch, M. E. (1983). The intensification of gender-related role expectations during early adolescence. In J. Brooks-Gunn & A. C. Peterson (Eds.), *Girls at puberty* (pp. 201–228). New York: Plenum Press.
- Hill, P. W., & Rowe, K. J. (1996). Multilevel modelling in school effectiveness research. *School Effectiveness and School Improvement*, 7, 1–34.
- Hox, J. (1995). *Applied multilevel analysis*. Amsterdam: TT Publikaties.
- Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, 60, 581–592.
- Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in children's self-competence and values: Gender and domain differences across grades one through twelve. *Child Development*, 73, 509–527.
- Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2001). Peer review in the funding of research in higher education: The Australian experience. *Educational Evaluation and Policy Analysis*, 23, 343–364.
- Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2002). A multilevel cross-classified modelling approach to peer review of grant proposals: The effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 166, 279–300.
- Jóhannesson, I. Á. (2004). To teach boys and girls: A pro-feminist perspective on the boys' debate in Iceland. *Educational Review*, 56, 33–42.
- Levesque, C., Zuehlke, A. N., Stanek, L. R., & Ryan, R. M. (2004). Autonomy and competence in German and American university students: A comparative study based on self-determination theory. *Journal of Educational Psychology*, 96, 68–84.
- Lingard, B., Martino, W., Mills, M., & Bahr, M. (2002). *Addressing the educational needs of boys*. Canberra, Australia: Department of Education, Science and Training.
- Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences*. Stanford, CA: Stanford University Press.
- MacDonald, A., Saunders, L., & Benfield, P. (1999). *Boys' achievement progress, motivation and participation: Issues raised by the recent literature*. Slough, England: National Foundation for Educational Research.
- Marsh, H. W. (1987a). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, 79, 280–295.
- Marsh, H. W. (1987b). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, 11, 253–388.
- Marsh, H. W. (1989a). Age and sex effects in multiple dimensions of self-concept: Preadolescence to early adulthood. *Journal of Educational Psychology*, 81, 417–430.
- Marsh, H. W. (1989b). Sex differences in the development of verbal and math constructs: The High School and Beyond study. *American Educational Research Journal*, 26, 191–225.
- Marsh, H. W. (1993). The multidimensional structure of academic self-concept: Invariance over gender and age. *American Educational Research Journal*, 30, 841–860.
- Marsh, H. W. (2007a). *Self-concept theory, measurement and research into practice: The role of self-concept in educational psychology*. Leicester, England: British Psychological Society.
- Marsh, H. W. (2007b). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–384). New York: Springer.
- Marsh, H. W., & Hau, K. (2003). Big-fish-little-pond effect on academic self-concept: A cross-cultural (26-country) test of the negative effects of academically selective schools. *American Psychologist*, 58, 364–376.
- Marsh, H. W., & Hau, K.-T. (2007). Applications of latent-variable models in educational psychology: The need for methodological-substantive synergies. *Contemporary Educational Psychology*, 32, 151–170.
- Marsh, H. W., Hau, K. T., Sung, R. Y. T., & Yu, C. W. (2007). Childhood obesity, gender, actual-ideal body image discrepancies, and physical self-concept in Hong Kong children: Cultural differences in the value of moderation. *Developmental Psychology*, 43, 647–662.
- Marsh, H. W., Martin, A., & Debus, R. (2001). Individual differences in verbal and math self-perceptions: One factor, two factors, or does it depend on the construct? In R. Riding & S. Rayner (Eds.), *Self perception: International perspectives on individual differences* (pp. 149–170). Westport, CT: Ablex.
- Marsh, H. W., Martin, A. J., & Hau, K. (2006). A multimethod perspective on self-concept research in educational psychology: A construct validity approach. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 441–456). Washington, DC: American Psychological Association.
- Marsh, H. W., & Rowe, K. J. (1996). The negative effects of school-average ability on academic self-concept—An application of multilevel modeling. *Australian Journal of Education*, 40, 65–87.
- Marsh, H. W., Rowe, K., & Martin, A. (2002). PhD students' evaluations of research supervision: Issues, complexities and challenges in a nationwide Australian experiment in benchmarking universities. *Journal of Higher Education*, 73, 313–348.
- Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, 76, 297–416.
- Martin, A. J. (2001). The Student Motivation Scale: A tool for measuring and enhancing motivation. *Australian Journal of Guidance and Counselling*, 11, 1–20.
- Martin, A. J. (2002). *Improving the educational outcomes of boys*. Canberra, Australia: Department of Education, Youth and Family Services.
- Martin, A. J. (2003a). Boys and motivation: Contrasts and comparisons with girls' approaches to schoolwork. *Australian Educational Researcher*, 30, 43–65.
- Martin, A. J. (2003b). Enhancing the educational outcomes of boys: Findings from the A. C. T. investigation into boys' education. *Youth Studies Australia*, 22, 27–36.
- Martin, A. J. (2003c). *How to motivate your child for school and beyond*. Sydney, Australia: Bantam.
- Martin, A. J. (2003d). The Student Motivation Scale: Further testing of an instrument that measures school students' motivation. *Australian Journal of Education*, 47, 88–106.
- Martin, A. J. (2004). School motivation of boys and girls: Differences of

- degree, differences of kind, or both? *Australian Journal of Psychology*, 56, 133–146.
- Martin, A. J. (2005). Exploring the effects of a youth enrichment program on academic motivation and engagement. *Social Psychology of Education*, 8, 179–206.
- Martin, A. J. (2006a). *The Motivation and Engagement Scale*. Sydney, Australia: Lifelong Achievement Group.
- Martin, A. J. (2006b). The relationship between teachers' perceptions of student motivation and engagement and teachers' enjoyment of and confidence in teaching. *Asia-Pacific Journal of Teacher Education*, 34, 73–93.
- Martin, A. J. (2007). Examining a multidimensional model of student motivation and engagement using a construct validation approach. *British Journal of Educational Psychology*, 77, 413–440.
- Martin, A. J. (in press). Enhancing student motivation and engagement: The effects of a multidimensional intervention. *Contemporary Educational Psychology*.
- Martin, A. J., & Marsh, H. W. (2005). Motivating boys and motivating girls: Does teacher gender really make a difference? *Australian Journal of Education*, 49, 320–334.
- Martin, A. J., & Marsh, H. W. (2006). Academic resilience and its psychological and educational correlates: A construct validity approach. *Psychology in the Schools*, 43, 267–281.
- Martin, A. J., Marsh, H. W., Williamson, A., & Debus, R. L. (2003). Self-handicapping, defensive pessimism, and goal orientation: A qualitative study of university students. *Journal of Educational Psychology*, 95, 617–628.
- Martino, W., & Meyenn, B. (2002). "War, guns and cool, tough things": Interrogating single-sex classes as a strategy for engaging boys in English. *Cambridge Journal of Education*, 32, 303–324.
- Marx, A., Fuhrer, U., & Hartig, T. (1999). Effects of classroom seating arrangements on children's question-asking. *Learning Environments Research*, 2, 249–263.
- Ntoumanis, N., & Vazou, S. (2005). Peer motivational climate in youth sport: Measurement development and validation. *Journal of Sport & Exercise Psychology*, 27, 432–455.
- Papaioannou, A. (1995). Differential perceptual and motivational patterns when different goals are adopted. *Journal of Sport & Exercise Psychology*, 17, 18–34.
- Papaioannou, A., Marsh, H. W., & Theodorakis, Y. (2004). A multilevel approach to motivational climate in physical education and sport settings: An individual or a group level construct? *Journal of Sport & Exercise Psychology*, 26, 90–118.
- Rasbash, J., Steele, F., Browne, W., & Prosser, B. (2004). *A user's guide to MLwiN Version 2.0*. London: Institute of Education, University of London.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics*, 16, 295–330.
- Rowe, K. J. (2000, November). Exploring "real" effects from evidence-based research in teacher and school effectiveness: The educational performance of males and females in school and tertiary education. Paper presented at the Educational Attainment and Labour Market Outcomes: Factors Affecting Boys and Their Status in Relation to Girls meeting, Melbourne, Australia.
- Rowe, K. J., & Rowe, K. S. (2002). *What matters most: Evidence-based findings of key factors affecting the educational experiences and outcomes for girls and boys throughout their primary and secondary schooling* (Supplementary submission to House of Representatives Standing Committee on Education and Training: Inquiry Into the Education of Boys). Canberra, Australia: House of Representatives Standing Committee on Education and Training.
- Seifriz, J. J., Duda, J. L., & Chi, L. (1992). The relationship of perceived motivational climate to intrinsic motivation and beliefs about success in basketball. *Journal of Sport & Exercise Psychology*, 14, 375–391.
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.
- Teven, J. J., & McCroskey, J. C. (1997). The relationship of perceived teacher caring with student learning and teacher evaluation. *Communication Education*, 46, 1–9.
- Weaver-Hightower, M. (2003). The "boy turn" in research on gender and education. *Review of Educational Research*, 73, 471–498.

Received December 21, 2006

Revision received June 17, 2007

Accepted August 21, 2007 ■



# A Multilevel Study of Predictors of Student Perceptions of School Climate: The Effect of Classroom-Level Factors

Christine W. Koth, Catherine P. Bradshaw, and Philip J. Leaf

Johns Hopkins Bloomberg School of Public Health and the Johns Hopkins Center for the Prevention of Youth Violence

A positive school climate is an important component of successful and effective schools and thus is often an aim of schoolwide initiatives. Climate has traditionally been conceptualized as a school-level factor and is often assumed to be related to other school-level factors (e.g., school size). The current study examines variation in perceptions of climate based on individual-, classroom-, and school-level factors to determine the influence of predictors at multiple levels. Data come from 2,468 5th graders from 37 public elementary schools. Two aspects of students' perception of school climate, order and discipline, and achievement motivation are examined. Multilevel analyses in hierarchical linear modeling indicate that individual-level factors (race and sex) accounted for the largest proportion of variance in perceptions of school climate. School-level factors (e.g., school size and faculty turnover) and several classroom-level factors (e.g., characteristics of the teacher, class size, and the concentration of students with behavior problems) were also significant predictors of perceptions of climate. These findings suggest that characteristics of the classroom environment are important to consider when aiming to improve school climate.

*Keywords:* school climate, multilevel analysis, classroom environment, behavioral problems

Positive school climate is recognized as an important component of successful and effective schools (Brand, Felner, Shim, Seitsinger, & Dumas, 2003; Kreft, 1993; Miller & Fredericks, 1990). It is defined as shared beliefs, values, and attitudes that shape interactions between students, teachers, and administrators and set the parameters of acceptable behavior and norms for the school (Emmons, Comer, & Haynes, 1996; Kuperminc, Leadbeater, Emmons, & Blatt, 1997). School climate is a product of social interactions among students and with teachers, is influenced by educational and social values, and has been shown to relate to social situations within classrooms and to the school as a whole. It has been linked to academic achievement and performance (Battistich, Solomon, Kim, Watson, & Schaps, 1995; Griffith, 1999); student misconduct, aggression, and behavioral problems (Battistich & Hom, 1997; Battistich, Solomon, Kim, Watson, & Schaps, 1995; Kuperminc, Leadbeater, & Blatt, 2001; Kuperminc et al., 1997; Loukas & Robinson, 2004; Shochet, Dadds, Ham, & Montague, 2006; Welsh, 2000; Wilson, 2004); adjustment problems (Kuperminc et al., 1997); and social and personal attitudes (Battistich et al., 1995).

This multidimensional construct has been examined from different theoretical and methodological perspectives. Prior research

has typically assessed teachers and school staff to investigate their perceptions of school organization and identify specific attributes that distinguish effective from ineffective schools (Stockard & Mayberry, 1992). Recently, there has been increased interest in students' perceptions of the school environment among educators, researchers, and policymakers (Brand et al., 2003; Griffith, 1995, 1999, 2000; Kuperminc et al., 2001, 1997; Van Horn, 2003; Verkuyten & Thijs, 2002; Vieno, Perkins, Smith, & Santinello, 2005; Welsh, 2000).

From a social cognitive perspective (Bandura, 2001; Rogers, 1951), people tend to react to experiences as they subjectively perceive them, not necessarily to how the experiences are objectively. Consequently, students' perceptions of the school environment likely have a significant impact on their behavior at school and thus are important potential targets for school improvement initiatives that aim to enhance achievement and reduce discipline problems (Haynes, Emmons, & Ben-Avie, 1997). Since the No Child Left Behind Act of 2001, two aspects of school climate—achievement and safety—have become central in schools' improvements. A wide range of interventions have been proposed to address climate, some of which are aimed at individuals and others of which are more focused on classrooms or the school level. However, the impact of interventions on achievement and safety may depend on the target of the intervention. Therefore, it is important to identify specific factors at different ecological levels (student, classroom, and school) that may influence students' perceptions of these two aspects of school climate.

## Measuring School Climate

School climate is multidimensional in nature, and an important issue is determining the appropriate unit of analysis: individual students or groups of students. Most previous research has con-

---

Christine W. Koth, Catherine P. Bradshaw, and Philip J. Leaf, Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, and Johns Hopkins Center for the Prevention of Youth Violence.

Support for this project comes from Centers for Disease Control and Prevention Grants R49/CCR318627 and 1U49CE000728 and National Institute of Mental Health Grant 1R01MH67948-1A1. We thank Elizabeth A. Stuart for providing consultation regarding the statistical analyses.

Correspondence concerning this article should be addressed to Christine W. Koth, 624 North Broadway, 8th Floor, Baltimore, MD 21205; E-mail: ckoth@jhsph.edu

ceptualized climate as a property of the school and analyzed it at the school level (see Anderson, 1982, for a review). Typically, an indicator of the climate is assessed and correlated with indicators of students' average performance, school characteristics, or student body composition (e.g., Brookover, Schweitzer, Schneider, Beady, Flood, & Weisenbaker, 1978; Halpin & Croft, 1963; Walberg, 1968; Walberg & Anderson, 1968). However, aggregating individual ratings data to form a single group-level indicator assumes little variation in the perception of different groups within the school (i.e., students, teachers, and administrators) and precludes investigation of diversity in perceptions of the climate.

Not all researchers view climate as an organizational indicator. Several studies have emerged documenting significant variation both within schools (likely attributable to individual-level factors) and between schools (likely attributable to school-level factors), thereby illustrating the importance of a multilevel approach (Battistich et al., 1995; Bevans, Bradshaw, Miech, & Leaf, 2007; Brand et al., 2003; Griffith, 1999, 2000; Philips, 1997; Rowan, Raudenbush, & Kang, 1991; Van Horn, 2003; Vieno et al., 2005). Specifically, student-level factors such as race (Battistich et al., 1995; Griffith, 2000; Kuperminc et al., 2001, 1997) and sex (Battistich et al., 1995; Griffith, 1999, 2000; Kuperminc et al., 2001, 1997; Verkuyten & Thijs, 2002; Welsh, 2000) have been shown to be significantly related to perceptions of school climate, with male and minority students tending to perceive the environment less favorably.

Commonly examined school-level predictors of school climate include structural aspects of the school, such as school size (Griffith, 2000; McNeely, Nonnemaker, & Blum, 2002; Welsh, 2000), student-teacher ratio (Griffith, 1995), and student mobility (Griffith, 2000). Aggregated indicators of student characteristics (e.g., socioeconomic status and ethnicity; Battistich et al., 1995; McNeely et al.; Vieno et al., 2005) and school type (public vs. private or urban vs. rural; Vieno et al., 2005) have also been linked with perceptions of school climate. However, relatively few studies have investigated factors at the classroom level in relation to perceptions of the overall school climate.

### Classroom-Level Predictors of School Climate

Classroom dynamics are complex and similar to school climate in that they involve the relationships and interactions between teachers and students, among students, and the perceptions, attitudes, and behaviors of students and teachers within the classroom (Montague & Rinaldi, 2001). It is likely that the climate of specific classrooms varies within a single school and that classroom management, class composition, and teacher characteristics may influence students' experiences. Research has suggested that teacher management style is related to the social structure of the class (Roland & Galloway, 2002). Teachers with practices that include emphasis on prosocial values and cooperation and teachers who were supportive have experienced improvements in positive student behavior and an increase in students' perception of connectedness (Solomon, Battistich, Kim, & Watson, 1996). Classroom variables that are more descriptive such as gender and ethnic composition and class size have also been investigated. Two studies of Dutch students incorporated these descriptive variables into their analyses but found no significant effects in relation to school satisfaction (Verkuyten & Thijs, 2002) and school adjust-

ment (van der Oord & Van Rossem, 2002). Similarly, teacher characteristics such as full- versus part-time status and work experience have also been investigated (van der Oord & Van Rossem, 2002) and linked with students' perceptions of climate.

Another potential classroom-level predictor of perceptions of school climate is the students' proximal exposure to deviant or aggressive behavior in the classroom. A study of 134 first-grade classrooms found considerable variability in the level of aggressive behavior across classrooms within schools and that the children's social behaviors varied as a function of the group norm (Stormshak et al., 1999). Furthermore, a growing number of studies have shown that groups of children with a high concentration of aggressive members affect both the behavior of the members and the dynamics of the group itself. In a study of first-grade classroom environments, classes with a higher proportion of students with past behavior problems also had significantly higher teacher ratings of shy behavior (Werthamer-Larsson, Kellern, & Wheeler, 1991). Research has also demonstrated that aggressive or deviant children shift the social norms, such that deviant behavior becomes socially acceptable among the members (Wright, Giammarino, & Parad, 1986). In addition to increasing the risk for behavior problems among classmates (Dishion, McCord, & Poulin, 1999; Dishion, Spracklen, Andrews, & Patterson, 1996; Patterson, Dishion, & Yoerger, 2000; Thornberry & Krohn, 1997), aggregating or clustering deviant youths within classrooms may have a proximal (Bronfenbrenner & Ceci, 1994) negative influence on the classroom environment and affect the students' overall perception of the school climate. Taken together, these findings suggest that the concentration of children with behavior problems may be an important classroom-level factor to consider when examining variation in children's perception of the school environment.

Although the notion that classroom characteristics influence students' overall perception of school climate seems reasonable, few studies have actually examined the various individual and classroom characteristics simultaneously (Malin & Linnakylae, 2001). The vast majority of multilevel research on school climate has used two-level models. For example, Verkuyten and Thijs (2002) found that factors at both the individual level (sex and minority status) and the class level (average motivation for academics and incidence of peer victimization) were related to students' overall perceptions of the school climate. Research using three-level multilevel modeling (individual, class, and school) is even rarer (e.g., Vieno et al., 2005). In a study of Italian school children, Vieno et al. (2005) found that 84% of the effect on climate was accounted for by individual-level factors (e.g., student's sex, age, socioeconomic status, parental monitoring, and control), whereas 11% was accounted for by class-level factors (e.g., democratic classroom culture) and 4% by school-level factors (e.g., school size, private vs. public, extent of extracurricular activities, and school resources). These findings suggest that attempts to identify the causes and consequences of school climate could benefit from examining potential predictors at multiple levels of the child's ecology (Griffith, 1999, 2000).

### Overview of the Current Study

The present study examined two distinct aspects of school climate (school safety and willingness to learn) to determine the



potential influence and relative contribution of factors at various levels (individual, classroom, and school). We investigated previously established associations between school climate and individual-level (e.g., race and sex) and school-level factors (e.g., school enrollment and students' average income level) as well as classroom-level factors. Specifically, we predicted that classroom indicators (e.g., class size and concentration of students with behavior problems) and teacher characteristics (e.g., advanced education and number of years teaching at the school) would have a proximal influence on students' perceptions of the school climate, above and beyond the influence of individual- and school-level factors. Using a multilevel framework, we were able to isolate the amount of variance associated with factors at the individual, classroom, and school levels.

## Method

Data for this study were collected as part of a large-scale study of a schoolwide behavior support program called Positive Behavioral Interventions and Supports. Thirty-seven Maryland public elementary schools from five school districts (rural and suburban) volunteered to participate in the study. Of the schools, 21 were randomized to the intervention condition and 16 were assigned to the comparison condition. The current study includes data from the first year of the trial only, and no significant intervention effects were observed on student reports of the school climate at this time point.

## Participants

The sample included 2,468 students in 120 non-special education fifth-grade classrooms within 37 elementary schools. The student sample was 48.8% female, and the racial and ethnic breakdown of students was 45.1% Caucasian, 34.5% African American, 3.2% American Indian, 2.2% Asian, and 15% "other" or "multiethnic." Of the 120 fifth-grade teachers included in the current analyses, 81.7% were female, 85.0% were Caucasian, 61.3% had been teaching at the school for 4 or fewer years, and 55.8% had education equivalent to or above a master's degree. The fifth-grade class sizes ranged from 11 to 31 students ( $M = 23.1$ ,  $SD = 3.7$ ). Total school enrollment ranged from 239 to 881 students ( $M = 488.4$ ,  $SD = 146.7$ ). The percentage of school-level (teaching) faculty turnover ranged from 0% to 40.0% ( $M = 13.1$ ,  $SD = 8.9$ ), and student mobility ranged from 5.7% to 47.9% ( $M = 24.5$ ,  $SD = 10.2$ ). The percentage of students receiving free or reduced-price meals ranged from 7.3% to 80.5% ( $M = 40.8$ ,  $SD = 20.0$ ).

## Measures

*Student (Level 1).* In the spring, all fifth-grade students at participating schools were asked to complete the elementary school version of the School Development Program School Climate Survey (Haynes, Emmons, & Ben-Avie, 2001). The School Climate Survey is a widely used (e.g., Kuperminc et al., 2001) and well-validated measure of students' perception of climate. Prior research by Haynes et al. (2001) has indicated that the measure has strong psychometric properties, including internal consistency and interrater reliability. The School Climate Survey assesses student

demographic information (e.g., sex and race) and consists of 53 statements regarding current school conditions, which are coded as "agree" or "disagree." Two subscales, Order and Discipline and Academic Motivation, were analyzed in the current study. Because of the study design, only fifth-grade students were asked to complete the School Climate Survey subscales; no data were collected from students in other grades. The survey subscales were group administered by trained project staff members, who provided a brief overview of the purpose of the survey and read each question aloud as the students completed the survey. The individual surveys were anonymous but were linked to the student's homeroom teacher.

The Order and Discipline subscale consists of 11 items (e.g., "My school is a safe place," "Children in my school fight a lot," and "At my school children disobey the rules") and assesses school safety and the appropriateness of student behavior at school. The Achievement Motivation subscale consists of six items (e.g., "My teachers believe I can do well in my school," "I feel I can do well in this school," and "I enjoy learning at this school") and assesses the extent to which the students believe they can and are willing to learn. Negatively stated items were reverse scored such that a higher score indicates a more positive school climate. The subscale scores were calculated by computing the percentage of items students agreed with on a scale ranging from 0 to 100. Analyses of the internal reliabilities of these two subscales indicated that the Order and Discipline subscale had a Cronbach's alpha of .74, and the Achievement Motivation subscale had an alpha of .63. Preliminary analysis using ordinary least squares (OLS) regression on individual student school climate scores and clustering on school (with robust standard errors) revealed no significant differences between intervention and comparison schools on either the Order and Discipline subscale ( $p = .39$ ) or the Achievement Motivation subscale ( $p = .85$ ).

*Teacher and classroom (Level 2).* Teachers completed a brief demographic questionnaire in the fall including questions regarding their gender, education, and number of years teaching at this school (with 4 years or less indicating newer teachers and 5 years or more indicating established teachers). Teacher ratings of individual students' disruptive or aggressive behaviors were obtained in the spring using the Teacher Observation of Classroom Adaptation—Checklist (Leaf, Schultz, Keys, & Ialongo, 2002). This checklist contains 25 items on a 6-point scale ranging from 1 (*never*) to 6 (*almost always*). The Aggressive/Disruptive subscale was used to calculate the percentage of students exhibiting behavior problems within each classroom and includes nine items (e.g., "breaks rules," "fights," "harms property," and "teases classmates"). The subscale scores were calculated by first averaging the nine items (Cronbach's  $\alpha = .93$ ) and then applying a cutoff to categorize students as displaying either adaptive behavior or problematic behavior. The cutoff was approximately 1 standard deviation above the total sample mean, such that 25.7% of the students across the full sample of participants were classified as exhibiting problematic behavior. Similar methods have been used to classify students in previous studies using versions of the Teacher Observation of Classroom Adaptation—Revised (August, Bloomquist, Lee, Realmuto, & Hektner, 2006; Petras, Chilcoat, Leaf, Ialongo, & Kellam, 2004; Schaeffer, Petras, Ialongo, Poduska, & Kellam, 2003; Lavalley, Bierman, & Nix, 2005; Stormshak et al., 1999). We then created a single concentration of students with behavior

problems score for each classroom by dividing the number of students classified with problematic behaviors in each class by the total number of students in the class. This procedure resulted in a mean of 26.0% ( $SD = 19.7$ ) of students per class being designated as exhibiting disruptive behavior.

*School characteristics (Level 3).* School enrollment, faculty turnover (percentage of faculty new to the school that year), student mobility (number of students migrating in plus the number migrating out, divided by total enrollment), and average student household income (percentage of students receiving free or reduced-price meals) were obtained from the Maryland State Department of Education for the school year. The receipt of free or subsidized lunches has been shown to be a good marker for low household income (Ensminger et al., 2000).

## Analyses

Preliminary descriptive analyses were conducted in STATA 9.2 (StataCorp, 2005) and indicated that the means for the Order and Discipline and Achievement Motivation subscales were 48.86 ( $SD = 22.94$ ) and 60.10 ( $SD = 27.24$ ), respectively. In addition, the two climate outcomes were correlated at .42. We used a multilevel approach to examine our main hypothesis that the clustering of students within classrooms accounts for a substantial portion of the variance in perceptions of school climate, above and beyond variation between students and the amount of variance accounted for clustering students within schools. Furthermore, we hypothesized that specific classroom-level factors, such as teacher characteristics and indicators of classroom disorder (e.g., large class size, high concentration of students with behavior problems) would be associated with student perceptions of school climate, even after controlling for individual- and school-level factors. We also explored possible within-level interactions separately for each level. A multilevel modeling technique was selected for the present study because both the data (students nested within classrooms nested within schools) and the hypotheses (the impact of school- and classroom-level factors on students' perceptions) are multi-level in nature (Raudenbush & Bryk, 2002). Single-level models are inappropriate for the current analyses because they assume that regression coefficients apply equally to all contexts (Duncan, Jones, & Moon, 1998; Luke, 2004). In addition, because individuals from the same school contexts will likely have correlated errors, a basic assumption of multivariate regression is violated (Luke, 2004). Multilevel modeling procedures account for non-independence of observations (students within classrooms, within schools) and allow for correlated error structures.

To examine the impact of students' perceptions within classrooms clustered within schools, we estimated three-level hierarchical linear models using HLM 6.02 software (Raudenbush, Bryk, & Congdon, 2005). All outcomes of school climate were measured at the student level (Level 1). Additional Level 1 indicators included individual student characteristics, Level 2 indicators included teacher and classroom variables, and Level 3 indicators included school characteristics. For each school climate outcome, an unconditional model with no covariates was estimated to partition the variance across the three levels. Two additional multilevel models were estimated for each outcome. First, Level 1 and Level 3 covariates were introduced to the model, then, to examine the influence of classroom-level factors above and beyond the

influence of other-level factors, the Level 2 covariates were added to the model with the Level 1 and Level 3 covariates. At each step of the model building, each parameter was inspected individually to assess the significance of the residual variance. Any covariates with nonsignificant variances were fixed (Hox, 1995; Raudenbush & Bryk, 2002). Model assumptions were carefully checked for each outcome. The assumption of homogeneity of residuals was tested by examining the normal probability plot of residual dispersion and the scatter plot of the Level 2 expected versus fitted scores (Luke, 2004; Raudenbush & Bryk, 2002). There was no evidence to suggest heteroscedasticity of the residuals. In addition, the possibility that the predicted values fell outside of the 0–100 range was examined. The data revealed that all values fell within that range. Maximum likelihood estimation with robust standard errors was used to estimate the parameters, and the overall fit of the models was evaluated on the basis of examination of the Akaike information criterion (Akaike, 1974) and the likelihood ratio test (Luke, 2004; Raudenbush & Bryk, 2002).

## Results

### Unconditional Model

Using HLM, we calculated the amount of variance for each of the three levels (student, classroom, and school) by fitting an unconditional model (without any covariates) for each school climate outcome (Raudenbush & Bryk, 2002). The partitioning of variance for each outcome is displayed in Table 1. The majority of the variance (65% for order and discipline and 86% for achievement motivation) was explained by between-student variation, and the clustering of students within schools accounted for an additional 5% to 27% of the variance in perceptions of achievement motivation and order and discipline, respectively. Whereas the majority of previous studies did not examine the clustering of students within classrooms, these analyses indicated that clustering at this level accounted for an additional 8%–9% of the total variance in student perceptions of order and discipline and achievement motivation, respectively. These findings illustrate the potential importance of considering variation on a classroom level.

### Multivariate Results

*Within-level interactions.* Using STATA, we conducted a series of single-level OLS regression analyses for each of the three levels of the hypothesized models to explore the possibility of within-level interactions to be included in the subsequent HLM multilevel models. Each OLS regression was clustered on schools

Table 1  
*Partitioning of Variance Across Levels From the Unconditional Multilevel Models for Students' Perceptions of Order and Discipline and Achievement Motivation*

Level	Order and discipline variance (%)	Achievement motivation variance (%)
1: Student	65	86
2: Classroom	8	9
3: School	27	5



to better estimate the standard errors. No significant interactions were found between the covariates within the Level 1 (student) or Level 3 (school) variables. Using the Level 2 data, an interaction between class size and number of years a teacher has been teaching at the school (newer teachers = 4 years or less and established teachers = 5 years or more) was detected. For ease of interpretation, the class size variable was centered at the grand mean ( $M = 23.1$ ). A single-level OLS regression indicated that this interaction was statistically significant for order and discipline ( $p = .04$ ) and achievement motivation ( $p = .03$ ). The calculated interaction term was entered as a separate Level 2 variable in the multilevel models to examine the association of the interaction within the multilevel framework.

**HLM.** Multilevel model estimates for order and discipline and achievement motivation are displayed separately in Table 2. Model

1 contains student-level (Level 1) and school-level (Level 3) covariates, and Model 2 contains estimates when the classroom (Level 2) covariates are added together with student- and school-level covariates. In Model 1, students' race and sex were significant for both of the school climate subscales. These associations remained in Model 2 when classroom-level covariates were added. These findings indicate that male and minority students perceived the school climate less favorably than did female and Caucasian students. Regarding the school-level factors in Model 1, the percentage of students from lower income households was statistically significant for order and discipline; however, when classroom factors were added to the model, this relation did not remain statistically significant.

We then included the classroom-level factors in the models to determine whether the individual- and school-level influences re-

Table 2  
*Multilevel Results for Order and Discipline and Achievement Motivation*

Level	Model 1			Model 2		
	Coefficient	SE	<i>t</i>	Coefficient	SE	<i>t</i>
Order and discipline						
1: Student						
Sex	-2.50**	0.87	-2.88	-2.43**	0.87	-2.79
Race	-3.93**	0.83	-4.74	-4.05**	0.80	-5.06
2: Classroom						
Class size				0.60†	0.32	1.86
Teaching years				3.64*	1.64	2.21
Class Size × Years				-0.98**	0.33	-3.00
Teacher education				-0.24	1.10	-0.22
% behavior problem				-0.27**	0.08	-3.40
3: School						
Enrollment	-0.02	0.01	-1.57	0.0	0.01	-0.61
Faculty turnover	-0.20	0.17	-1.20	-0.25**	0.08	-2.98
Student mobility	-0.05	0.16	-0.34	-0.08	0.10	-0.85
FARMS	-0.25**	0.08	-2.99	-0.08	0.05	-1.59
AIC	21,667.0			21,639.3		
Δ parameters	7					
Δ -2LL	41.63**					
Achievement motivation						
1: Student						
Sex	-6.66**	1.04	-6.41	-6.62**	1.04	-6.37
Race	-2.80*	1.29	-2.17	-2.74*	1.24	-2.20
2: Classroom						
Class size				0.92**	0.36	2.56
Teaching years				2.01	2.17	0.92
Class Size × Years				-1.08*	0.47	-2.32
Teacher education				-1.36	2.22	-0.62
% behavior problem				-0.12*	0.05	-2.19
3: School						
Enrollment	-0.02†	0.01	-1.94	-0.02**	0.01	-2.64
Faculty turnover	0.05	0.11	0.45	0.14	0.10	1.45
Student mobility	0.03	0.14	0.23	-0.06	0.09	-0.61
FARMS	-0.05	0.06	-0.86	0.01	0.05	0.22
AIC	23,067.7			23,063.3		
Δ parameters	7					
Δ -2LL	18.41**					

*Note.* For sex, 1 = male, 0 = female; for race, 1 = minority, 0 = Caucasian; class size = number of students in the class; teaching years = number of years teacher has taught at this school; for teacher education, 1 = master's level or higher, 0 = bachelor's degree; % behavior problem = percentage of students in the class with behavior problems; enrollment = number of students enrolled at the school; FARMS = free or reduced price meals; AIC = Akaike information criterion.

\*\*  $p \leq .01$ . \*  $p \leq .05$ . †  $p \leq .10$ .

mained significant (see Model 2 in Table 2). There were no significant changes in individual-level factors when compared with Model 1. Regarding the school-level factors, faculty turnover was related to order and discipline, such that students from schools with higher turnover described the school as less orderly. With regard to school enrollment, we found that larger schools tended to be associated with lower scores on achievement motivation in Model 1, and this association became statistically significant in Model 2 when accounting for classroom-level factors.

Focusing on the classroom-level factors (see Model 2 of Table 2), we found the percentage of disruptive students in a class to be negatively associated with both indicators of school climate. Students in classrooms with a greater number of disruptive students rated the school climate less favorably than did students in classrooms with fewer disruptive peers. The within-Level 2 interaction (class size and number of years teaching at the school) remained significant for order and discipline and achievement motivation in the multilevel analyses. This interaction term represents the difference in school climate as class size increases among students of more established teachers compared with students of newer teachers. For both order and discipline and achievement motivation, the interaction coefficient was negative, indicating that in larger size classes, the students of more established teachers perceived the school climate less favorably than did students of newer teachers. The coefficient for class size was positive and statistically significant for achievement motivation and represents the change in school climate scores of students of newer teachers in larger versus smaller classes; therefore, for students with newer teachers, a larger class size was associated with a more positive academic setting than was a smaller class size. Because class size was mean-centered, the coefficient for years teaching represents the difference in school climate scores between established and newer teachers with an average class size of 23 students and was statistically significant for order and discipline. In other words, students of more established teachers in average-sized classrooms reported a more orderly climate than students of newer teachers in average-sized classrooms.

**Model fit.** A series of fit indices were calculated to evaluate the fit of the data to the final models. As shown in Table 2 for the two school climate variables, the models with student-, classroom-, and school-level covariates had lower Akaike information criteria (indicating better fit) and a significant difference in the likelihood ratio test ( $p < .01$ ) as compared with the models with only student- and school-level covariates.

## Discussion

The present study used a multilevel framework to examine the influence of individual-, classroom-, and school-level factors on students' perceptions of school climate. When we unpack these influences across levels in the unconditional model, we see that the largest proportion of variance comes from individual-level factors (65%–86%; Table 1). Further inspection of proportion of variance across the three levels for each of the climate outcomes indicates that achievement motivation had the lowest amount of school-level variance (5%) and the highest amount of individual-level variance (86%). The proportion of variance suggests there is greater variability in students' willingness to learn within schools and that this aspect of school climate may be more indicative of individuals'

own motivation than is overall aggregated perception. In contrast, the amount of school-level variance for order and discipline was much higher (27%), suggesting that perceptions of school safety may be more relevant to school characteristics than achievement motivation; however, the individual level still accounted for the majority of the variance. Last, 8% to 9% of the variance across the two climate outcomes was attributable to clustering at the classroom level. This partitioning of variance is relatively consistent with previous research by Vieno et al. (2005), who found that 84% of the variation in climate was accounted for at the individual level, whereas 11% was accounted for at the class level, and just 4% at the school level.

From a methodological perspective, these findings suggest that researchers should pay careful attention to the clustering of students, both within schools and within classes, when examining school climate in intervention trials or cross-sectional epidemiological studies (Luke, 2004). Overlooking the nesting of students would likely increase the Type I error rate; therefore, researchers should adjust the standard errors to obtain accurate estimates (Murray, 1998).

## Individual-Level Factors

Consistent with previous research (Battistich et al., 1995; Griffith, 1999, 2000; Kuperminc et al., 2001, 1997; Verkuyten & Thijs, 2002; Welsh, 2000), individual-level factors such as race and sex were associated with perceptions of the school environment, with male and minority students tending to perceive the school less favorably. Male students reported less order and discipline and lower levels of achievement motivation even after controlling for school- and classroom-level factors in Model 2. Prior research has indicated that boys are more likely than girls to display disruptive behavior at school (Lahey et al., 2000; McDermott, 1996; Putallaz & Bierman, 2004; Roberts & Baird, 1972; Tremblay et al., 1996) and therefore may perceive the environment as less safe and orderly. With regard to achievement motivation, boys tend to receive lower grades in elementary school than girls, which may contribute to this difference in their willingness to learn. With regard to race, minority students perceived the environment as less safe and reported lower levels of achievement motivation than did Caucasian youths, even after controlling for classroom- and school-level factors. These findings may reflect cultural differences in the expectations in the school setting (Zimmerman, Khoury, Vega, Gil, & Warheit, 1995) or in the construct validity of school climate (Kuperminc et al., 1997). Interventions that aim to increase a sense of positive climate should raise mutual understanding and awareness of culturally linked expectations in schools.

## School-Level Factors

A series of school-level factors, including school size, faculty turnover, student mobility, and student free or reduced-price meals rate, were also examined as predictors of student perceptions of the school environment. These effects were smaller than we had anticipated. Specifically, larger enrollment was significantly negatively associated with achievement motivation, and high faculty turnover was related to lower perceptions of order and discipline, after controlling for influences at the other two levels. Although



prior research suggests that elementary school students can begin to feel lost or disconnected in large schools (Hellman & Beaton, 1986), Griffith (2000) did not find school size to be a significant factor in students' perceptions of school climate. These discrepant findings suggest that school size may be related to some aspects of school climate but not to others, and additional research with a larger sample of more diverse schools may further clarify the potential association between variables typically considered to be school-level indicators of school disorder (Birnbaum et al., 2003) and students' perceptions of school climate.

### *Classroom-Level Factors*

As hypothesized, some of the classroom-level variables were associated with students' perceptions of the school environment. Of particular interest is the impact of clusters of students with behavior problems. As expected, the greater the proportion of students with behavior problems in a classroom, the less favorably the students perceived the school environment. This effect was strongest for perceptions of order and discipline. These findings suggest that children are particularly sensitive to their classmates' behavior problems and that students perceive the school's safety and their willingness to learn in relation to the number of disruptive classmates.

We detected a significant interaction effect between class size and number of years teaching for both outcomes. Regarding order and discipline, in an average class size of 23, students with more established teachers perceived school climate as safer than did students with newer teachers. Given just this information, one might theorize that teachers who have been working at the school for several years are better integrated into the school, able to provide a more stable and predictable environment for the students, and more familiar with the students. However, the significant interaction term indicates this is not true for all students and teachers. Specifically, students in larger classes with more established teachers tended to view the school environment as less safe than did students in smaller classes with more established teachers. In addition, students in larger classes with newer teachers perceived the school environment as safer than those in smaller classes with newer teachers. Regarding achievement motivation, there was no statistical difference in students' willingness to learn between newer and more established teachers of average class size. However, students in larger classes of newer teachers were more willing to learn than students in smaller classes of newer teachers. In addition, students' reports of their willingness to learn in larger classrooms of more established teachers were less favorable than those in smaller classrooms.

This interaction is an intriguing finding. It is important to remember that the students' perceptions are in regards to overall school climate and not classroom climate, and the factors that contribute to the dynamics of the classroom may affect school climate differently. Students are not randomly assigned to classrooms or teachers; therefore, it is possible that this interaction is indirectly measuring some other construct. Perhaps our measurement of the number of years teaching (i.e., newer vs. more experienced teachers) is a proxy measure for teaching styles or teacher-student interactions that were not measured directly. It is also possible that newer teachers in our study were exposed to different

teaching methods while receiving their education than were their more experienced colleagues.

Other factors could explain why students of different classroom compositions perceive school climate differently. Studies suggest that teachers' perceptions and attitudes toward students affect their own behavior as well as students' behavior within the classroom (Ladd, Birch, & Buhs, 1999; Weinstein, Madison, & Kuklinski, 1995). Osterman (2000) contended that students with a sense of belonging and positive involvement in the classroom are more likely to demonstrate acceptance of authority and regulate their own behavior in the classroom. In addition, teacher interactions with students seem to influence students' perceptions of one another (Birch & Ladd, 1997). Prior research has indicated that teachers who display favoritism or are perceived as not being fair to all students can negatively influence the sense of community (Altenbaugh, Engel, & Martin, 1995). Students' attitudes about their teacher also tend to influence their sense of school satisfaction, such that school satisfaction is higher for students who like their teacher and have a more supportive relationship with the teacher (Baker, 1999; Verkuyten & Thijs, 2002). The current study was not designed to examine process-oriented variables, such as teaching style and teacher-student interaction; therefore, additional research on these types of variables within a multilevel framework is needed. It should also be noted that examining within-level interactions within the multilevel framework is rare. Methodologically, further exploration into these interactions is needed to investigate the validity of the relationship we uncovered.

### *General Discussion*

The results of the current study indicate that student- and classroom-level factors tend to have greater influence on students' perceptions of the school environment than do school-level factors. Interventions aiming to enhance students' perceptions may be most effective if they target those with the most negative attitudes, such as male and minority students. There are several individual-level factors that were not examined in the current study that might also influence students' perceptions, such as their academic abilities, social relationships, socioeconomic status, and own problem behavior. Future research should examine these factors more specifically as possible predictors of school climate that may help target individual interventions more effectively. With regard to school-level factors, there are several initiatives focused on creating smaller schools and learning environments, and our findings suggest that school size was only marginally inversely related to climate. Efforts to increase the connectedness of within-school groupings, such as improving relations between teachers and students and those between peers within classrooms, may have a more favorable impact on students' perceptions of school climate than focusing on efforts to affect school-level factors (e.g., reducing school size). Reducing class size has often been cited as a possible strategy for increasing academic performance; however, our findings, along with those of other climate studies, suggest that class size alone may not greatly influence perceptions of school climate (Griffith, 1995; van der Oord & Van Rossem, 2002; Verkuyten & Thijs, 2002). Taken together, these findings suggest that factors at several levels should be assessed when examining different aspects of school climate and developing initiatives to enhance school climate.



## References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723.
- Altenbaugh, R. J., Engel, D. E., & Martin, D. T. (1995). *Caring for kids: A critical study of urban school leaders*. Bristol, PA: Falmer.
- Anderson, C. S. (1982). The search for school climate: A review of the research. *Review of Educational Research*, 53, 368–420.
- August, G. J., Bloomquist, M. L., Lee, S. S., Realmuto, G. M., & Hektner, J. M. (2006). Can evidence-based prevention programs be sustained in community practice settings? The early risers' advanced-stage effectiveness trial. *Prevention Science*, 7, 151–165.
- Baker, J. A. (1999). Teacher-student interaction in urban at-risk classrooms: Differential behavior, relationship quality, and student satisfaction with school. *Elementary School Journal*, 100, 57–70.
- Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology*, 52, 1–26.
- Battistich, V., & Hom, A. (1997). The relationship between students' sense of their school as a community and their involvement in problem behaviors. *American Journal of Public Health*, 87, 1997–2001.
- Battistich, V., Solomon, D., Kim, D., Watson, M., & Schaps, E. (1995). Schools as communities, poverty levels of student populations, and students' attitudes, motives, and performance: A multilevel analysis. *American Educational Research Journal*, 32, 627–658.
- Bevans, K. B., Bradshaw, C. P., Miech, R., & Leaf, P. J. (2007). Staff- and school-level predictors of school organizational health: A multilevel analysis. *Journal of School Health*, 77, 294–303.
- Birch, S. H., & Ladd, G. W. (1997). The teacher-child relationship and children's early school adjustment. *Journal of School Psychology*, 35, 61–80.
- Birnbaum, A. S., Lytle, L. A., Hannan, P. J., Murray, D. M., Perry, C. L., & Forester, J. L. (2003). School functioning and violent behavior among young adolescents: A contextual analysis. *Health Education Research*, 18, 389–403.
- Brand, S., Felner, R., Shim, M., Seitsinger, A., & Dumas, T. (2003). Middle school improvement and reform: Development and validation of a school-level assessment of climate, cultural pluralism, and school safety. *Journal of Educational Psychology*, 95, 570–588.
- Bronfenbrenner, U., & Ceci, S. (1994). Nature–nurture reconceptualized in developmental perspective: A bioecological model. *Psychological Review*, 101, 568–586.
- Brookover, W. B., Schweitzer, J. H., Schneider, J. M., Beady, C. H., Flood, P. K., & Weisenbaker, J. M. (1978). Elementary school social climate and school achievement. *American Educational Research Journal*, 15, 301–318.
- Dishion, T. J., McCord, J., & Poulin, F. (1999). When interventions harm: Peer groups and problem behavior. *American Psychologist*, 54, 755–764.
- Dishion, T. J., Spracklen, K. M., Andrews, D. W., & Patterson, G. R. (1996). Deviancy training in male adolescent friendships. *Behavior Therapy*, 27, 373–390.
- Duncan, C., Jones, K., & Moon, G. (1998). Context, composition and heterogeneity: Using multilevel models in health research. *Social Science and Medicine*, 46, 97–117.
- Emmons, C. L., Comer, J. P., & Haynes, N. M. (1996). Translating theory into practice: Comer's theory of school reform. In J. P. Comer, N. M. Haynes, E. Joyner, & M. Ben-Avie (Eds.), *Rallying the whole village* (pp. 27–41). New York: Teachers College Press.
- Ensminger, M. E., Forrest, C. B., Riley, A. W., Kang, M., Green, B. F., Startfield, B., & Ryan, S. A. (2000). The validity of measures of socioeconomic status of adolescents. *Journal of Adolescent Research*, 15, 392–419.
- Griffith, J. (1995). An empirical examination of a model of social climate in elementary school. *Basic and Applied Social Psychology*, 17, 97–117.
- Griffith, J. (1999). School climate as “social order” and “social action”: A multi-level analysis of public elementary school student perceptions. *School Psychology of Education*, 2, 339–369.
- Griffith, J. (2000). School climate as group evaluation and group consensus: Student and parent perceptions of the elementary school environment. *Elementary School Journal*, 101, 35–61.
- Halpin, A. W., & Croft, D. B. (1963). *The organizational climates of schools*. Chicago: University of Chicago, Midwest Administration Center.
- Haynes, N. M., Emmons, C. L., & Ben-Avie, M. (1997). School climate as a factor in student adjustment and achievement. *Journal of Educational and Psychological Consultation*, 8, 321–329.
- Haynes, N. M., Emmons, C. L., & Ben-Avie, M. (2001). *The School Development Program Student, Staff, and Parent School Climate Surveys*. New Haven, CT: Yale Study Center.
- Hellman, D. A., & Beaton, S. (1986). The pattern of violence in urban public schools: The influence of school and community. *Journal of Research in Crime and Delinquency*, 23, 102–127.
- Hox, J. J. (1995). *Applied multilevel analysis*. Amsterdam: TT-Publikaties.
- Kreft, I. G. G. (1993). Using multilevel analyses to assess school effectiveness: A study of Dutch secondary school. *Sociology of Education*, 66, 104–129.
- Kuperminc, G. P., Leadbeater, B. J., & Blatt, S. J. (2001). School social climate and individual differences in vulnerability to psychopathology among middle school students. *Journal of School Psychology*, 39, 141–159.
- Kuperminc, G. P., Leadbeater, B. J., Emmons, C., & Blatt, S. J. (1997). Perceived school climate and difficulties in the social adjustment of middle school students. *Applied Developmental Science*, 1, 76–88.
- Ladd, G. W., Birch, S. H., & Buhs, E. S. (1999). Children's social and scholastic lives in kindergarten: Related spheres of influence. *Child Development*, 70, 1373–1400.
- Lahey, B. B., Schwab-Stone, M., Goodman, S. H., Waldman, I. D., Canino, G., Rathouz, P. J., Miller, T. L., Dennis, K. D., Bird, H. R., & Jensen, P. S. (2000). Age and gender differences in oppositional behavior and conduct problems: A cross-sectional household study of middle childhood and adolescence. *Journal of Abnormal Psychology*, 109, 488–503.
- Lavallee, K. L., Bierman, K. L., & Nix, R. L. (2005). The impact of first-grade “friendship group” experiences on child social outcomes in the fast track program. *Journal of Abnormal Child Psychology*, 33, 307–324.
- Leaf, P. J., Schultz, D., Keys, S., & Ialongo, N. (2002). *The Teacher Observation of Classroom Adaptation—Checklist (TOCA-C)*. Baltimore: Johns Hopkins Center for the Prevention of Youth Violence.
- Loukas, A., & Robinson, S. (2004). Examining the moderating role of perceived school climate in early adolescent adjustment. *Journal of Research on Adolescence*, 14, 209–233.
- Luke, D. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage.
- Malin, A., & Linnakylae, P. (2001). Multilevel modeling in repeated measures of the quality of Finnish school life. *Scandinavian Journal of Educational Research*, 45, 145–166.
- McDermott, P. A. (1996). A nationwide study of developmental and gender prevalence for psychopathology in children and adolescence. *Journal of Abnormal Child Psychology*, 24, 53–66.
- McNeely, C. A., Nonnemaker, J. M., & Blum, R. W. (2002). Promoting school connectedness: Evidence from the National Longitudinal Study of Adolescent Health. *Journal of School Health*, 72, 138–146.
- Miller, S. I., & Fredericks, J. (1990). The false ontology of school climate effects. *Educational Theory*, 40, 333–342.
- Montague, M., & Rinaldi, C. (2001). Classroom dynamics and children at risk: A follow-up. *Learning Disability Quarterly*, 24, 75–83.
- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.
- No Child Left Behind Act of 2001, Pub. L. 107–110, 115 Stat. 1425 (2002).



- Osterman, K. F. (2000). Students' need for belonging in the school community. *Review of Educational Research*, 70, 323-367.
- Patterson, G. R., Dishion, T. J., & Yoerger, K. (2000). Adolescent growth in new forms of problem behavior: Macro- and micro-peer dynamics. *Prevention Science*, 1, 3-13.
- Petras, H., Chilcoat, H. D., Leaf, P. J., Jalongo, N. S., & Kellam, S. G. (2004). Utility of TOCA-R scores during the elementary school years in identifying later violence among adolescent males. *Journal of the American Academy of Child & Adolescent Psychiatry*, 43, 88-96.
- Philips, M. (1997). What makes schools effective? A comparison of the relationship of communitarian climate and academic climate to mathematics achievement and attendance during middle school. *American Educational Research Journal*, 34, 633-662.
- Putallaz, M., & Bierman, K. L. (2004). *Aggression, antisocial behavior, and violence among girls: A developmental perspective*. New York: Guilford Press.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Raudenbush, S. W., Bryk, A., & Congdon, R. (2005). *HLM 6*. Lincolnwood, IL: Scientific Software International.
- Roberts, J., & Baird, J. T. (1972). *Behavior patterns of children in school. Vital and health statistics* (DHEW Pub. No. HSM 72-1042). Washington, DC: U.S. Government Printing Office.
- Rogers, C. R. (1951). *Client-centered therapy*. Boston: Houghton-Mifflin.
- Roland, E., & Galloway, D. (2002). Classroom influences on bullying. *Educational Research*, 44, 299-312.
- Rowan, B., Raudenbush, S. W., & Kang, S. J. (1991). Organizational design in high schools: A multilevel analysis. *American Journal of Education*, 99, 238-266.
- Schaeffer, C. M., Petras, H., Jalongo, N., Poduska, J., & Kellam, S. (2003). Modeling growth in boys' aggressive behavior across elementary school: Links to later criminal involvement, conduct disorder, and antisocial personality disorder. *Developmental Psychology*, 39, 1020-1035.
- Shochet, I. M., Dadds, M. R., Ham, D., & Montague, R. (2006). School connectedness is an underemphasized parameter in adolescent mental health: Results of a community prediction study. *Journal of Clinical Child and Adolescent Psychology*, 35, 170-179.
- Solomon, D., Battistich, V., Kim, D., & Watson, M. (1996). Teacher practices associated with students' sense of the classroom as a community. *Social Psychology of Education*, 1, 235-267.
- StataCorp. (2005). *Stata 9*. College Station, TX: StataCorp LP.
- Stockard, J., & Mayberry, M. (1992). *Effective educational environments*. Newbury Park, CA: Corwin Press.
- Stormshak, E. A., Bierman, K. L., Bruschi, C., Dodge, K. A., Coie, J. D., & the Conduct Problems Prevention Research Group. (1999). The relation between behavior problems and peer preference in different classroom contexts. *Child Development*, 70, 169-182.
- Thornberry, T. P., & Krohn, M. D. (1997). Peers, drug use, and delinquency. In D. M. Stoff, J. Breiling, & J. D. Maser (Eds.), *Handbook of antisocial behavior* (pp. 218-233). New York: Wiley.
- Tremblay, R. E., Boulerice, B., Harden, P. W., McDuff, P., Perusse, D., Pihl, R. O., & Zoccolillo, M. (1996). Do children in Canada become more aggressive as they approach adolescence? In *Growing up in Canada* (pp. 127-137). Ottawa: Statistics Canada.
- van der Oord, E. J., & Van Rossem, R. (2002). Differences in first graders' school adjustment: The role of classroom characteristics and social structure of the group. *Journal of School Psychology*, 40, 371-394.
- Van Horn, M. L. (2003). Assessing the unit of measurement for school climate through psychometric and outcome analyses of the school climate survey. *Educational and Psychological Measurement*, 63, 1002-1019.
- Verkuyten, M., & Thijs, J. (2002). School satisfaction of elementary school children: The role of performance, peer relations, ethnicity and gender. *Social Indicators Research*, 59, 203-228.
- Vieno, A., Perkins, D. D., Smith, T. M., & Santinello, M. (2005). Democratic school climate and sense of community in school: A multilevel analysis. *American Journal of Community Psychology*, 36, 327-341.
- Walberg, H. J. (1968). Structural and affective aspects of classroom climate. *Psychology in the Schools*, 5, 247-253. (ERIC Reproduction Service Document No. ED015154)
- Walberg, H. J., & Anderson, G. J. (1968). Classroom climate and individual learning. *Journal of Educational Psychology*, 59, 414-419.
- Weinstein, R. S., Madison, S., & Kuklinski, M. (1995). Raising expectations in schooling: Obstacles and opportunities for change. *American Educational Research Journal*, 32, 121-159.
- Welsh, W. N. (2000). The effect of school climate on school disorder. *Annals of the American Academy of Political and Social Sciences*, 567, 88-107.
- Werthamer-Larsson, L., Kellam, S. G., & Wheeler, L. (1991). Effect of first-grade classroom environment on child shy behavior, aggressive behavior, and concentration problems. *American Journal of Community Psychology*, 19, 585-602.
- Wilson, D. (2004). The interface of school climate and school connectedness and relationships with aggression and victimization. *Journal of School Health*, 74, 293-299.
- Wright, J. C., Giammarino, M., & Parad, H. W. (1986). Social status in small groups: Individual-group similarity and the social "misfit." *Journal of Personality and Social Psychology*, 50, 523-536.
- Zimmerman, R. S., Khoury, E. L., Vega, W. A., Gil, A. G., & Warheit, G. J. (1995). Teacher and parent perceptions of behavior problems among a sample of African American, Hispanic and non-Hispanic White students. *American Journal of Community Psychology*, 23, 181-196.

Received October 3, 2006

Revision received August 21, 2007

Accepted August 30, 2007 ■

# The Role of Achievement Goals in the Development of Interest: Reciprocal Relations Between Achievement Goals, Interest, and Performance

Judith M. Harackiewicz  
University of Wisconsin—Madison

Amanda M. Durik  
Northern Illinois University

Kenneth E. Barron  
James Madison University

Lisa Linnenbrink-Garcia  
Duke University

John M. Tauer  
University of St. Thomas

The dynamics of individual and situational interest and academic performance were examined in the college classroom and 7 semesters later in conjunction with achievement goals. At the beginning of an introductory psychology course, participants reported their initial interest in psychology, achievement goals, and situational interest in course lectures. At the end of the semester, participants ( $N = 858$ ) reported their situational interest in course lectures and psychology. In the short term, relationships emerged among initial interest, achievement goals, situational interest, and class performance. Longitudinally, situational interest during the introductory course, independent of initial interest, predicted subsequent course choices. Results are discussed in terms of S. Hidi and K.A. Renninger's (2006) 4-phase model of interest development and the multiple goals model (J. M. Harackiewicz, K. E. Barron, P. R. Pintrich, A. J. Elliot, & T. M. Thrash, 2002).

**Keywords:** achievement goals, interest, achievement motivation, academic motivation, academic performance

One of the critical developmental tasks that students face in college is to identify and solidify their interests as they select courses, choose their academic major, and make decisions about career paths. Why do some students become involved and interested in their coursework, and why do they continue in a particular academic discipline? Do students learn more and perform better in their classes when they are interested? These questions highlight the importance of interest in college education, both as an important outcome variable in its own right (Harackiewicz, Barron, & Elliot, 1998; Hidi & Harackiewicz, 2000; Maehr, 1976) and as a motivational factor that may influence learning and performance (Hidi, 1990; Hidi & Renninger, 2006).

Students' achievement goals may play an important role in shaping academic motivation and interest because they reflect the

purpose of achievement behavior in a particular setting and can influence how a student approaches coursework (Dweck & Leggett, 1988; Nicholls, 1984). When pursuing mastery goals in a learning situation, students want to develop competence by acquiring new knowledge and skills. When pursuing performance-approach goals, students want to demonstrate competence relative to others, and when pursuing performance-avoidance goals, they hope to avoid the demonstration of incompetence (Elliot & Church, 1997; Middleton & Midgley, 1997). In addition to goals focused on competence, some students adopt work-avoidance goals that focus on effort minimization (Brophy, 1983; Nicholls, 1989).

Because students can and do pursue multiple goals in their classes, researchers have begun to examine the independent and interactive effects of these goals on performance and motivation (Harackiewicz et al., 1998; Pintrich, 2000). Although there is little debate about the negative effects of performance-avoidance and work-avoidance goals (e.g., Elliot & Church, 1997; Harackiewicz, Barron, Carter, Lehto, & Elliot, 1997; Middleton & Midgley, 1997), the pattern of mastery and performance-approach goal effects has been mixed on measures of interest and performance (see Harackiewicz, Barron, Pintrich, Elliot, & Thrash, 2002, for review). Several studies have found positive effects of performance-approach goals, but not mastery goals, on grades in high school and college classes (Archer, 1994; Bouffard, Boisvert, Vezeau, & Larouche, 1995; Bouffard, Vezeau, & Bordeleau, 1998; Elliot & Church, 1997). In contrast, several studies have found

---

Judith M. Harackiewicz, Department of Psychology, University of Wisconsin—Madison; Amanda M. Durik, Department of Psychology, Northern Illinois University; Kenneth E. Barron, Department of Psychology, James Madison University; Lisa Linnenbrink-Garcia, Department of Psychology and Neuroscience, Duke University; John M. Tauer, Department of Psychology, University of St. Thomas.

We thank Jonathan Trinastic for his help with this research, and we thank Mary Ainley, Suzanne Hidi, and Ann Renninger for their insightful comments on drafts of this article.

Correspondence concerning this article should be addressed to Judith M. Harackiewicz, Department of Psychology, University of Wisconsin—Madison, 1202 West Johnson Street, Madison, WI 53705. E-mail: jmharack@wisc.edu



positive effects of mastery goals, but not performance goals, on interest in classes. For example, Harackiewicz et al. (1997) found that performance-approach goals predicted higher grades in an introductory psychology class, whereas mastery goals predicted interest in the class.

Harackiewicz, Barron, Tauer, and Elliot (2002) replicated these results and extended the study of goal effects over time by following introductory psychology students until they graduated. Performance-approach goals predicted grades in an introductory psychology class, grades in subsequent psychology classes, and academic GPA over the ensuing semesters. The performance-approach goal effects remained significant even with prior levels of academic performance and standardized measures of ability controlled. In contrast, mastery goals predicted interest in the class, continued interest in psychology, and whether students chose to major in psychology. Considered together, this pattern of findings supports a multiple goal perspective (Barron & Harackiewicz, 2000; Harackiewicz, Barron, Pintrich, et al., 2002; Pintrich, 2000) in which mastery and performance-approach goals can both promote important, but distinct, educational outcomes.

### Reciprocal Effects of Mastery Goals and Interest

The positive associations between mastery goals and interest are well documented (Harackiewicz, Barron, Pintrich, et al., 2002), but the direction of causality has not been examined closely. Previous longitudinal research has tested whether the adoption of mastery goals early in a course is a predictor of subsequent interest (whether measured later in the course or over the course of an undergraduate career) but has not tested whether initial levels of interest (measured at the beginning of a course) predict the adoption of mastery goals. It is important to consider the level of interest in place before students formulate their goals for a course. Students who begin a class with higher levels of initial interest may be more likely to adopt mastery goals because they want to learn more about the domain that interests them. In turn, these mastery goals may promote the continued development of interest, suggesting that mastery goals might actually mediate interest processes. In other words, initial interest might predispose students to adopt mastery goals, which then predict subsequent interest. Although previous goals research has considered mastery goals as predictors of interest (e.g., Harackiewicz et al., 1997; Harackiewicz, Barron, Tauer, Carter, & Elliot, 2000; Harackiewicz, Barron, Tauer, & Elliot, 2002), the reverse direction of causality seems equally plausible: Interest in the course content might predict the adoption of mastery goals because students who are interested in a topic may want to learn more about it (Renninger, 2000).

Another possibility is that initial interest predicts both mastery goal adoption and subsequent interest, and is therefore a third variable that accounts for the mastery goal-subsequent interest relationship. Similar concerns about reverse causality and third-variable explanations for the relationship between performance-approach goals and course grades have been addressed in previous research by including measures of prior academic performance and standardized ability tests (e.g., Harackiewicz, Barron, Tauer, & Elliot, 2002), but these issues have not been addressed with respect to mastery goal effects on interest. Only by measuring initial interest before goal adoption will it be possible to disentangle

some of these intriguing possibilities and explore the reciprocal effects of mastery goals and interest.

### The Development of Interest

A careful examination of interest development requires consideration of the interests and experience that students bring to class, as well as the types of interest that can develop or deepen in classes over time. Some students may enter a course with a high level of interest in the topic because they have had prior experience with related material and have found it interesting. These students' interest might deepen and solidify over the course of a semester. Other students may begin a course with limited knowledge and little initial interest in the topic but discover or develop interest during the course depending on features of the course, such as whether the instructor or course materials stimulate their attention and involvement. This highlights the two different types of interest discussed in the literature. Interest that resides within the individual over time has been distinguished from interest that emerges in response to situational cues. The first type has been labeled *individual interest*; it has a dispositional quality and is deep and enduring. Renninger (2000) argued that individual interest requires having substantial knowledge of a topic as well as valuing that knowledge. Students who enter a course with a high level of initial interest in the topic may be described as having an individual interest. The second type has been labeled *situational interest*; it emerges spontaneously in response to features of the environment (Dewey, 1913; Hidi, 1990).

Hidi and Baird (1986) raised an important distinction between two types of situational interest, reflecting the process through which situational interest is first triggered and then maintained in a particular situation. Situational factors such as instructional style or collative variables can first trigger or "catch" interest through attentional and affective processes, but once situational interest is aroused, it may or may not last, depending on the external supports in the situation (Hidi & Renninger, 2006; Mitchell, 1993). In a college classroom, for example, an entertaining lecture might stimulate students' attention and catch their interest, but course material that is personally meaningful or valued may better hold students' interest (Harackiewicz et al., 2000; Malone & Lepper, 1987; Rathunde, 1993; Schiefele, 1991; Wigfield & Eccles, 1992). If classroom factors promote the development of meaning and value, situational interest may be maintained over time. If this maintained interest endures beyond the particular situation and is associated with the accumulation of knowledge about the topic, it may eventually become a deep, individual interest (Hidi & Renninger, 2006; Krapp, 2002). Therefore, it is also important to distinguish two types of situational interest: "caught" interest (associated with arousal, attention, and affect) and "held" interest (associated with personal value and meaning).

Our research concerns the interest (both individual and situational) that can develop or deepen over the course of an introductory college course, taking initial levels of interest into account. Students might come to an introductory class with a well-developed individual interest (indeed, they may have chosen to take that class because of their prior knowledge and valuing of the course content), and their individual interest may deepen if they find value or personal meaning in the course material. In contrast, for students who begin the class without strong interest, the de-



velopment of interest must begin with situational interest. Because situational interest is externally supported, students who begin a class without much interest in the domain will only develop interest if features of the class trigger or catch their situational interest (Hidi & Renninger, 2006). This situational interest may be maintained or held if students come to find meaning in the course material.

The experiences students have during an introductory course may promote continuing interest that has the potential to develop into a deeper individual interest. For example, students who enter a course with high initial interest may become motivated to take more courses in that content area not only because of their ongoing individual interest, but also because they have come to appreciate the content even more as a consequence of taking the course. Students who begin a course with low initial interest may also develop continuing interest and pursue further coursework in the content area if their situational interest can be held over time. Held situational interest may eventually develop into individual interest if it is accompanied by deepening knowledge and sustained over time and across contexts (Hidi & Renninger, 2006).

### The Current Research

In the present research we examined the development of interest in an academic domain using an achievement goal paradigm. We studied a sample of students in an introductory psychology course, allowing us to examine their initial interest, background knowledge, and goals in a foundational course. This design extends previous research in several ways, most notably by including multiple measures of interest (both individual and situational) and by situating these measures in a temporal sequence relative to the assessment of achievement goals and in relation to students' receipt of performance feedback in the course. Previous research in college courses has only measured situational interest late in the semester, after students have received extensive performance feedback, making it impossible to disentangle interest-performance relations. In the present study we included an early measure of situational interest ("catch-1"), administered before any exams were given, allowing us to test situational interest as an uncontaminated predictor of performance. In addition, we obtained students' grades on the first exam in the course, which allowed us to examine the effects of early performance on the subsequent development of situational interest (both "catch-2" and "hold") measured later in the course. Thus, in addition to measuring initial interest, the current study included a measure of situational interest that was uncontaminated by performance feedback, as well as an early indicator of performance, allowing us to test the reciprocal relations between interest and performance across the semester.

We had three primary objectives for the current research. First, we wanted to integrate goal theory (Pintrich, 2002) with interest theory (Hidi & Renninger, 2006) by examining the role of individual interest and mastery goals in promoting the development of situational and continuing interest in an academic domain. We hypothesized that initial interest would directly predict students' adoption of mastery goals, situational interest, and continued interest. Some students are drawn to courses because they have initial interest in the topic. Once in the course, these students should want to learn more about the topic (adopt mastery goals) because interest directly influences the desire to gain knowledge

about the domain (Renninger, 2000). The course should further expose these individuals to the content so that their initial interest may continue to deepen. However, we also hypothesized that mastery goals would promote students' situational and continuing interest independent of initial interest. A mastery goal orientation may lead students to approach their coursework in a task-focused manner that allows them to become excited about and involved in the learning experience. Mastery goals may therefore be particularly effective in facilitating the triggering of situational interest, which when combined with personal meaning and/or value may develop further into a held situational interest. Thus, individual interest and mastery goals may represent related but relatively distinct paths to the development of situational and continued interest.

Our second objective was to examine the reciprocal relations between interest and performance in a classroom context. Hidi (1990) has argued that interest can promote performance, but in previous classroom studies, interest measures have only been collected late in the course, after students have received considerable performance feedback (e.g., Harackiewicz et al., 1997), making it difficult to test this hypothesis. In the current study, we collected a measure of situational interest before any exams were administered in the course, enabling us to examine the effect of situational interest on later performance. Also, because we obtained an early measure of performance and later measures of situational interest, it was possible to test the effects of performance on later interest. Our design therefore allowed us to test both possibilities simultaneously. Accordingly, we hypothesized that students who reported situational interest would perform better in the class, and we also predicted that students who performed well on early exams would develop interest in the class.

Our third objective was to examine the predictors of goal adoption. We have already discussed our prediction that initial interest should predict mastery goal adoption, but it is important to compare initial interest with other variables that have been shown to predict achievement goal adoption in previous research. Harackiewicz et al. (1997; Harackiewicz, Barron, Tauer, & Elliot, 2002) found that students who were high in workmastery, one component of achievement motivation identified by Spence and Helmreich (1983), were particularly likely to adopt mastery goals and reject work-avoidance goals, whereas students high in competitiveness (a second component of achievement motivation) were more likely to adopt performance-approach goals. Of particular interest in the current study was whether this pattern would be replicated when we controlled for the effects of initial interest. If both initial interest and workmastery predict students' adoption of mastery goals, then we will have identified two independent predictors of mastery goals, one that taps an individual's response to the specific course content (initial interest) and one that taps an individual's characteristic response to achievement situations (workmastery).

We situated the present analyses of interest processes and goal adoption within the context of our previous goals research. We expected to replicate both the short-term and long-term goal effects documented by Harackiewicz et al. (1997, 2000; Harackiewicz, Barron, Tauer, & Elliot, 2002) and Elliot and Church (1997). Specifically, we hypothesized that performance-approach goals would positively predict grades; that work-avoidance and performance-avoidance goals would negatively predict grades; and that mastery goals and initial interest would both predict situa-



tional interest, continued interest over the following seven semesters, and whether students chose to major in psychology. Finally, we hypothesized that the effects of goals and initial interest on long-term outcomes would be mediated through their impact on short-term outcomes. Specifically, we expected mastery goals and initial interest to predict continued interest in psychology and majoring in psychology because they promote the development of held situational interest in the topics covered in the introductory course.

## Method

### Overview

In a prospective, longitudinal study, we collected data at five points in time from students enrolled in introductory psychology courses at a large midwestern university. The first four assessments were during the semester that students were enrolled in Introductory Psychology, and for the fifth assessment, we obtained data from students' academic records seven semesters later. During the first week of the semester (Time 1), we measured students' initial interest in psychology, as well as individual differences in achievement orientation. During the third week of the semester (Time 2), before any exams or feedback, we measured students' achievement goals for Introductory Psychology as well as their enjoyment of the course at that point (catch-1). Two weeks before the end of the semester (Time 3), we measured students' enjoyment of the class (catch-2) and their valuation of the course material (hold). Upon completion of the semester (Time 4), we obtained students' final grade in Introductory Psychology as well as their overall semester GPA. Finally, seven semesters after the completion of the introductory course (Time 5), we obtained students' transcripts and coded them for continued interest in psychology, majoring in psychology, grades in subsequent psychology classes, and overall GPA.

### Participants and Setting

We recruited students from five classes of Introductory Psychology (approximately 430 students per class). Only students who were taking the course for graded credit while pursuing an undergraduate degree were included. This excluded, for example, undergraduate students who were taking the course for pass/fail credit as well as graduate students. Classes were almost entirely lecture format. Students took three or four multiple-choice exams (the number of exams varied across classes), and final grades were based on a normative curve recommended by the psychology department.

Data collection at Time 1 was part of a large departmental survey (1,265 participants) for which Introductory Psychology students received extra credit in exchange for their participation. Participants enrolled in our study at Time 2, when they signed consent forms and authorized access to their academic records. Students completed questionnaires during class time and did not receive extra credit for participation in the remainder of the study. Of the original 1,265, 1,128 participants entered the study at Time 2. After enrollment in the study, some attrition occurred over subsequent assessments. First, 27 participants dropped the course before the semester ended (course drops). Second, because data

were collected during class time, students absent from class at Time 3 ( $n = 206$ ) were not included in the study (absentee drops). Finally, 37 participants took fewer than 15 additional credits over the following seven semesters (equivalent to one full-time semester), which precluded coding of long-term follow-up measures (university drops). Thus, the final sample (used in all data analyses unless otherwise noted) included 858 students (279 men, 579 women) and represented 76% of the students who enrolled in our study at Time 2.

### Time 1: Initial Interest and Achievement Orientation

During the first week of the semester, students completed a questionnaire about their interest and background in psychology. All items were rated on a 7-point scale (1 = *not at all true of me*, 7 = *very true of me*). Seven items assessed initial interest in psychology, and three items measured students' previous experience in psychology (background knowledge). We based items on those used by Barron and Harackiewicz (2003) and wrote new items to represent Renninger's (1992) conceptualization of individual interest. Items and reliabilities for all scales are presented in the Appendix.

Participants also completed the Work and Family Orientation Questionnaire (Helmreich & Spence, 1978; Spence & Helmreich, 1983). This personality measure assesses individual differences in three components of achievement motivation: work orientation, mastery orientation, and competitiveness. Each item was rated on a 5-point scale (1 = *strongly disagree*, 5 = *strongly agree*). The work and mastery items were combined into a single index as recommended by Spence and Helmreich, resulting in a two-dimensional measure of achievement orientation: workmastery and competitiveness. These scales have been shown to have good reliability and validity (Helmreich & Spence, 1978; Spence, Pred, & Helmreich, 1989).

### Time 2: Achievement Goals and Situational Interest

Three weeks into the semester, students completed questions about their achievement and work-avoidance goals. All items were rated on a 7-point scale (1 = *not at all true of me*, 7 = *very true of me*) and referred to their current goal orientation in Introductory Psychology. Four different types of goals were assessed: mastery, performance-approach, performance-avoidance, and work-avoidance. Scales were adapted from those used by Harackiewicz et al. (1997), Elliot and Church (1997), Midgley et al. (1996), and Pintrich (Pintrich & DeGroot, 1990; Pintrich & Garcia, 1991).

After completing questions about their goal orientations, students were asked to focus on their reactions to the class and to report their situational interest (catch-1). Participants responded to questions about their experiences in and enjoyment of course lectures on a 7-point scale (1 = *strongly disagree*, 7 = *strongly agree*). Items were adapted from Harackiewicz et al. (2000; Harackiewicz, Barron, Tauer, & Elliot, 2002) to assess the catch component of situational interest, based on Hidi's theoretical work (Hidi & Baird, 1986; Hidi & Berndorff, 1998).

### Time 3: Interest

Near the end of the semester (13 weeks into the term, with only 2 weeks remaining), students responded to questions about their

interest in the course. All items were rated on a 7-point scale (1 = *strongly disagree*, 7 = *strongly agree*). Five items assessed students' affective reactions to course lectures (catch-2), similar to the measure at Time 2. Nine other items assessed students' feelings about and personal valuing of the course material (hold; Krapp, 2002; Schiefele, 1991). Linnenbrink-Garcia et al. (2007) conducted exploratory and confirmatory factor analyses with these items and demonstrated that catch and hold represented separate components of situational interest.

#### *Time 4: Performance in Introductory Psychology and Semester GPA*

We obtained measures of students' performance in Introductory Psychology, which consisted of students' scores on the first exam in the course (first exam score), as well as their overall final grade (final grade). Because different scales were used by each instructor for the first exam, we standardized first exam scores within each section of the course. In contrast, final grades were assigned using the same letter-grade scale for all sections and were distributed similarly within sections. Based on a 4-point scale ( $A = 4.0$ ,  $F = 0.0$ ), the mean grade for students in our study was 2.99 ( $SD = 0.79$ ). Grades were distributed as follows:  $A = 22.6\%$ ,  $AB = 19.5\%$ ,  $B = 19.3\%$ ,  $BC = 16.3\%$ ,  $C = 19.6\%$ ,  $D = 2.7\%$ ,  $F = 0\%$ . We also obtained students' overall semester GPA for the semester in which they took Introductory Psychology.

#### *Time 5: Continued Interest and Subsequent Performance*

Seven semesters later, we obtained students' academic records and used them to compute four long-term outcome measures. First, we counted the number of additional psychology course credits students had taken over the years, providing a behavioral measure of continued interest in psychology (psychology courses taken). The mean number of additional psychology credits taken was 5.23 ( $SD = 10.32$ ), with the typical psychology course at this university worth 3 credits. We also recorded whether students had declared a major yet and, if so, whether they had majored in psychology (psychology major).

We also calculated two long-term performance measures. To measure academic performance in subsequent psychology courses, we computed a psychology GPA for those students who had taken additional psychology classes (psychology GPA). Finally, as a general measure of academic performance, we also computed a GPA for all courses taken after the semester of the initial study (subsequent GPA).

## Results

### *Attrition Analyses*

Three sets of attrition analyses were conducted comparing the final sample to (a) participants who dropped the course (course drops;  $n = 27$ ), (b) participants who were not in class at Time 3 (absentee drops;  $n = 206$ ), and (c) participants who completed the study but did not enroll in enough subsequent courses for inclusion in analyses on long-term follow-up measures (university drops;  $n = 37$ ). Comparisons were made on all variables allowed by the patterns of missing data for each group.

The group of students who dropped the course (course drops) were compared with the final sample on all Time 1 and Time 2 variables. Independent sample  $t$  tests revealed four significant differences. Effect sizes (Cohen's  $d$ ) were computed for these differences, revealing moderate effects. The final sample had higher scores on workmastery,  $t(883) = 2.00$ ,  $p < .05$ ,  $d = .39$ ; and mastery goals,  $t(883) = 2.56$ ,  $p < .05$ ,  $d = .50$ ; and lower scores on performance-avoidance goals,  $t(883) = -2.31$ ,  $p < .05$ ,  $d = -.45$ ; and work-avoidance goals,  $t(883) = -2.84$ ,  $p < .01$ ,  $d = -.56$ . No differences emerged for initial interest, background knowledge, competitiveness, performance-approach goals, or catch-1.

We then compared students who were absent for Time 3 (absentee drops) with the final sample on all Time 1 and Time 2 variables and on final grade. Independent sample  $t$  tests revealed five significant differences. There were small effects showing that the final sample was higher on initial interest,  $t(1062) = 2.34$ ,  $p < .05$ ,  $d = .18$ ; workmastery,  $t(1062) = 3.12$ ,  $p < .01$ ,  $d = .24$ ; and mastery goals,  $t(1062) = 2.51$ ,  $p < .05$ ,  $d = .19$ ; and lower on work-avoidance goals,  $t(1062) = -2.89$ ,  $p < .01$ ,  $d = -.22$ . The final sample also had higher final grades, revealing a moderate effect,  $t(1062) = 7.53$ ,  $p < .01$ ,  $d = .59$ . No reliable differences emerged for background knowledge, competitiveness, performance-approach goals, performance-avoidance goals, or catch-1.

Finally, we compared the final sample to the group of university drops on all variables from Times 1 to 4. Two significant differences emerged. The final sample received higher grades,  $t(893) = 3.10$ ,  $p < .01$ ,  $d = .52$ , and was lower in work-avoidance goals,  $t(893) = -1.96$ ,  $p = .05$ ,  $d = -.33$ , than the university drops. Considered together, these results suggest that students who expressed initial interest, adopted mastery goals, or achieved higher grades were somewhat overrepresented in our sample, whereas students who adopted performance-avoidance and work-avoidance goals were somewhat underrepresented.

### *Descriptive Statistics and Factor Analyses*

The means and standard deviations for all variables are reported in Table 1 along with the zero-order intercorrelations. Although the structure, content, and grading distributions of the five class sections were comparable, we tested for instructor differences in all variables. There were significant instructor effects on several variables, indicating that students reported higher levels of some goals and types of interest in some sections compared with others. Therefore, we included four dummy code terms (D1–D4) to test and control for mean-level differences between the instructors in all subsequent analyses. One class was arbitrarily designated as the control group (Cohen & Cohen, 1983).

We conducted an exploratory factor analysis of our goal and interest measures that were collected at Time 2 to document the difference between goals and situational interest. There were 20 items, and we used principal components extraction with oblimin rotation to determine their structure. There were five eigenvalues greater than 1.0, accounting for 70% of the total item variance. Moreover, the pattern matrix evidenced simple structure that corresponded to the five predicted factors (catch-1, mastery goal, performance-approach goal, performance-avoidance goal, work-avoidance goal). Items loaded on their intended factors (none



Table 1  
Zero-Order Correlations and Descriptive Statistics

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. Workmastery	—																		
2. Competitiveness	.25*	—																	
3. Individual interest	.18*	-.11*	—																
4. Background knowledge	.04	-.04	.21*	—															
5. Mastery goals	.33*	-.11*	.56*	.12*	—														
6. Performance-approach goals	.01	.36*	-.01	-.07	-.04	—													
7. Performance-avoidance goals	-.19*	.14*	-.19*	-.04	-.28*	.19*	—												
8. Work-avoidance goals	-.30*	.16*	-.32*	-.11*	-.46*	.15*	.34*	—											
9. Catch-1	.18*	-.10*	.32*	.05	.44*	-.02	-.22*	-.31*	—										
10. First exam grade	.09*	.06	.06	.00	.06	.21*	-.08	-.04	.12*	—									
11. Catch-2	.12*	-.10*	.28*	.02	.32*	-.01	-.18*	-.23*	.62*	.22*	—								
12. Hold	.16*	-.11*	.60*	.17*	.54*	-.02	-.22*	-.32*	.42*	.20*	.60*	—							
13. Final grade in psychology	.10*	.04	.05	-.02	.07	.18*	-.11*	-.08	.10*	.82*	.20*	.19*	—						
14. Semester GPA	.08	.04	.02	-.03	.04	.17*	-.08	-.09*	.06	.65*	.09*	.09*	.77*	—					
15. Courses taken	.04	-.03	.30*	.19*	.24*	.04	-.13*	-.18*	.14*	.09	.18*	.32*	.10*	.04	—				
16. Psychology GPA ( $N = 341$ )	.09	-.04	.09	-.06	.13	.18*	-.05	-.10	.07	.57*	.09	.12*	.62*	.60*	.13*	—			
17. Subsequent GPA	.06	-.03	.01	-.05	.02	.11*	-.10*	-.13*	.06	.47*	.05	.05	.56*	.64*	.03	.73*	—		
18. Major ( $N = 770$ )	.03	-.04	.27*	.19*	.20*	.03	-.11*	-.15*	.13*	.08	.16*	.30*	.07	.03	.92*	.09	.00	—	
19. Gender	-.07	.18*	-.24*	-.12*	-.19*	.06	.01	.25*	-.11*	-.04	-.15*	-.22*	-.07	-.07	-.16*	-.08	-.11*	-.12*	—
<i>M</i>	3.77	3.68	5.53	4.00	5.32	5.31	4.92	2.98	4.35	0.00	4.11	5.02	2.99	3.18	5.23	3.08	3.26	0.11	
<i>SD</i>	0.50	0.73	1.07	2.26	0.88	1.24	1.45	1.35	1.19	1.00	1.39	1.21	0.79	0.53	10.32	0.72	0.46	0.32	
Cronbach's alpha	.78	.75	.90	.90	.87	.87	.78	.90	.85		.91	.95							

Note.  $N = 858$  unless otherwise noted in parentheses.

\*  $p < .01$ .

lower than .57) and did not load on their unintended factors (none higher than .27).

We also conducted a factor analysis of all of our interest and goal measures (41 items) across Times 1 through 3 using the same procedure. We found that there were eight eigenvalues greater than 1.0, accounting for 70% of the variance. The pattern matrix again revealed simple structure, with the eight factors corresponding to our eight predicted factors (the five factors in the previous analysis with the addition of individual interest, catch-2, and hold), with each item loading on its intended factor (no loadings lower than .52) and not loading on unintended factors (no loadings higher than .30). These factor analyses indicated that the items were measuring separate constructs.

### Regression Models

We ran three sets of regression analyses, each predicting outcomes measured at different time points. We conducted one series of multiple regression analyses to examine the predictors of the achievement goals adopted in the class (Time 1 variables predicting Time 2 variables), a second series to examine the predictors of short-term outcome variables measured during the semester in which students took Introductory Psychology (Time 1 and 2 variables predicting Time 3 and 4 variables), and a final series to examine the predictors of long-term outcome variables collected in the follow-up assessment (Time 1, 2, 3, and 4 variables predicting Time 5 variables).

This data analytic strategy allowed us to retain the continuous nature of the variables and to test the independent effects of each predictor variable as well as the interactions between them. For the following analyses, all continuous main effect terms were standardized, and multiplicative two- and three-way interaction terms were created with these variables (Aiken & West, 1991). Interactions that were significant on any measure were retained in all models, but nonsignificant interactions were trimmed from models. We also retained the Mastery  $\times$  Performance-Approach Goal interaction in all models, because this represents one important test of the multiple goal perspective (Barron & Harackiewicz, 2001). To interpret significant interaction effects from these analyses, we computed predicted values ( $\hat{Y}$ s) for representative high and low

groups (1 *SD* above and below the mean) from the regression equations using the unstandardized regression coefficients. Because of the large sample size, statistical power was high. Therefore, we set alpha at .01 and only interpreted standardized regression coefficients that were .10 or greater.

### Predictors of Achievement Goals

Our first series of multiple regression analyses examined the effects of initial interest and achievement orientation on goals adopted in the classroom. Preliminary regression models included the main effect terms for initial interest, background knowledge, workmastery, competitiveness, gender (coded +1 for women and -1 for men), the four instructor dummy codes, as well as all two- and three-way interactions among these variables. No interactions were significant, but the Initial Interest  $\times$  Background interaction was retained for consistency with other models, resulting in a 10-term goals model. Table 2 summarizes each of the models we tested.

**Mastery goals.** The overall model was significant,  $F(10, 847) = 54.56, p < .001 (R^2 = .39)$ . Main effects were found for initial interest,  $F(1, 847) = 274.88, p < .001 (\beta = .48)$ ; workmastery,  $F(1, 847) = 90.13, p < .001 (\beta = .27)$ ; and competitiveness,  $F(1, 847) = 17.62, p < .001 (\beta = -.12)$ . Students high in initial interest, those high in workmastery, or those low in competitiveness were more likely to adopt mastery goals. In addition, one of the instructor dummy code contrasts (D3) was significant,  $F(1, 847) = 7.89, p < .01 (\beta = .10)$ , indicating a significant difference between two classes in mastery goal adoption.

**Performance-approach goals.** The overall model was significant,  $F(10, 847) = 16.38, p < .001 (R^2 = .16)$ . The main effect of competitiveness was significant,  $F(1, 847) = 131.54, p < .001 (\beta = .38)$ , and showed that students high in competitiveness were more likely to adopt performance-approach goals. In addition, one of the instructor dummy code contrasts (D2) was significant,  $F(1, 847) = 7.06, p < .01 (\beta = -.11)$ , indicating a difference between two classes in the extent to which students adopted performance-approach goals.

Table 2  
Regression Models Tested (Predictor Variables Included in Each Model)

Goals model (10 terms)
Initial interest; background; Initial Interest $\times$ Background interaction; workmastery; competitiveness; gender; dummy codes (4)
Dependent variables: mastery, performance-approach, performance-avoidance, and work-avoidance goals
Basic short-term model (15 terms)
Goals model with the addition of mastery, performance-approach, performance-avoidance, and work-avoidance goals; Mastery $\times$ Performance-Approach interaction
Dependent variables: catch-1; catch-2; hold; final grade
Catch-1 short-term model (16 terms)
Basic short-term model with the addition of catch-1
Dependent variables: catch-2; hold; final grade
Direct effects long-term model (16 terms)
Basic short-term model with the addition of number of credits taken
Dependent variables: psychology credits taken; psychology major; psychology GPA
Final long-term model (21 terms)
Direct effects long-term model with the addition of catch-1; catch-2; hold; final grade; and Hold $\times$ Grade interaction
Dependent variables: psychology credits taken; psychology major; psychology GPA

Note. The number of terms that were retained in each specific model after preliminary testing of higher order interactions appear in parentheses.



**Performance-avoidance goals.** The overall model was significant,  $F(10, 847) = 9.20, p < .001$  ( $R^2 = .10$ ); and the main effects of initial interest,  $F(1, 847) = 17.58, p < .001$  ( $\beta = -.15$ ); workmastery,  $F(1, 847) = 37.65, p < .001$  ( $\beta = -.21$ ); and competitiveness,  $F(1, 847) = 29.45, p < .001$  ( $\beta = .19$ ), were significant. Students high in initial interest or workmastery were less likely to endorse performance-avoidance goals. In contrast, students high in competitiveness were more likely to endorse performance-avoidance goals.

**Work-avoidance goals.** The overall model was significant,  $F(10, 847) = 28.14, p < .001$  ( $R^2 = .25$ ); and the main effects of initial interest,  $F(1, 847) = 33.69$  ( $\beta = -.19$ ); workmastery,  $F(1, 847) = 85.53, p < .001$  ( $\beta = -.29$ ),  $p < .001$ ; competitiveness,  $F(1, 847) = 33.49, p < .01$  ( $\beta = .18$ ); and gender,  $F(1, 847) = 19.98, p < .001$  ( $\beta = -.14$ ), were significant. Students high in initial interest or workmastery were less likely to endorse work-avoidance goals. Competitive students and male students were more likely to endorse work-avoidance goals. In addition, three of the instructor dummy code contrasts (D2, D3, and D4) were significant,  $F(1, 847) = 19.78, p < .001$  ( $\beta = -.17$ );  $F(1, 847) = 10.23, p < .01$  ( $\beta = -.12$ ); and  $F(1, 847) = 9.97, p < .01$  ( $\beta = -.13$ ), respectively, indicating that students in certain classes were more likely to adopt work-avoidance goals than those in other classes. Figure 1 depicts the significant paths from Time 1 variables to all four goals.

### Predictors of Short-Term Outcomes

**Regression model for short-term analyses.** In our second series of multiple regression analyses, we tested the short-term effects of initial interest, achievement orientation, and achievement goals on catch-1, catch-2, hold, and final grade. Preliminary models included the main effects of initial interest; background knowledge; workmastery; competitiveness; gender; the four in-

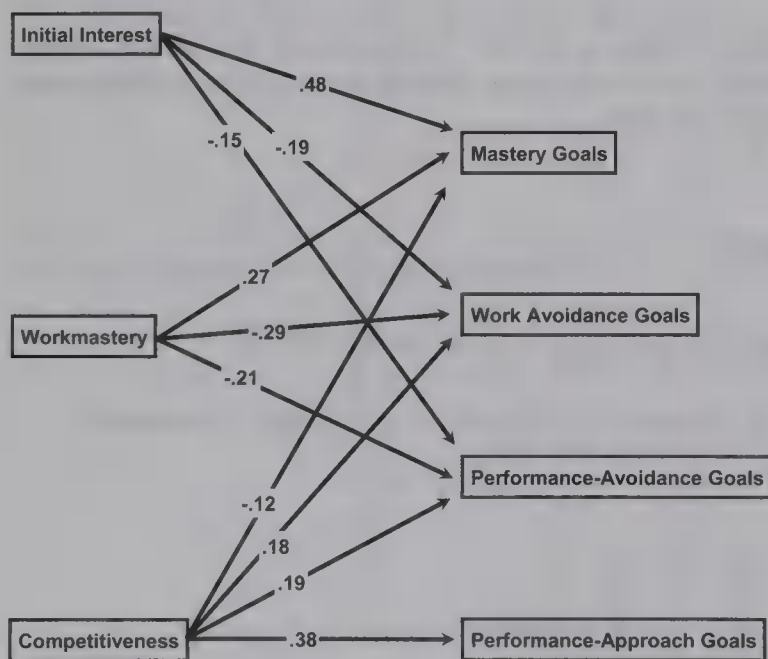


Figure 1. Predictors of achievement goals. All paths represented are significant ( $p < .01$ ). Path coefficients are standardized regression coefficients. For clarity of presentation, significant gender and instructor effects are not depicted, but these variables were controlled in these analyses.

structor dummy codes; mastery, performance-approach, performance-avoidance, and work-avoidance goals, as well as all two- and three-way interactions among these variables. Preliminary analyses revealed no significant higher order interactions, and thus the only interaction terms retained were those between initial interest and background knowledge, and between mastery and performance-approach goals. We therefore tested a 15-term basic short-term model (see Table 2).

**Catch-1.** The basic short-term model was significant,  $F(15, 842) = 21.08, p < .001$  ( $R^2 = .27$ ). A significant main effect was found for mastery goals,  $F(1, 842) = 53.06, p < .001$  ( $\beta = .29$ ), such that students who strongly endorsed mastery goals reported higher levels of interest in the lectures at Time 2 than students who did not endorse mastery goals. A significant main effect for work-avoidance goals,  $F(1, 842) = 10.70, p < .01$  ( $\beta = -.12$ ), showed that students who endorsed work-avoidance goals reported less interest in the lectures. A significant effect of one of the instructor codes (D3),  $F(1, 842) = 21.20, p < .001$  ( $\beta = .18$ ), indicated a difference between two sections in the extent to which students reported catch-1. Figure 2 depicts a path model of these effects.

We tested the predictors of catch-2, hold, and final grade in two steps. We first tested the basic 15-term model described previously, and then, because catch-1 was measured 2 months earlier than these three short-term outcome variables, we were also able to test for the effects of early situational interest. We therefore tested a second model in which catch-1 was included as a predictor of the later variables. Preliminary testing revealed no significant interactions of catch-1 with the terms included in the basic model, and thus the catch-1 short-term model included 16 terms comprising the basic model plus the inclusion of catch-1 (see Table 2). Figure 2 depicts the paths to the three short-term variables (catch-2, hold, and final grade) estimated from the catch-1 short-term model.

**Catch-2.** The basic short-term model was significant,  $F(15, 842) = 15.05, p < .001$  ( $R^2 = .21$ ). A significant main effect was found for mastery goals,  $F(1, 842) = 33.64, p < .001$  ( $\beta = .24$ ), such that students who adopted mastery goals reported higher levels of interest in the lectures at Time 3 than students who did not endorse mastery goals. In addition, two of the instructor dummy code terms (D3 and D4) were significant,  $F(1, 842) = 33.09, p < .001$  ( $\beta = .24$ ), and  $F(1, 842) = 7.03, p < .01$  ( $\beta = .10$ ), indicating that students in certain classes were more likely to report higher levels of catch-2 than those in other classes.

The catch-1 short-term model accounted for significantly more variance in catch-2 than the basic model ( $R^2_{\text{change}} = .21, p < .001$ ). This change was attributable to the significant effect of catch-1,  $F(1, 841) = 296.70, p < .001$  ( $\beta = .53$ ), suggesting that students who reported catch at Time 2 were more likely to report catch at Time 3 (see Figure 2). The effect of mastery goals was no longer significant ( $\beta = .09$ ). Following the procedures outlined by Kenny, Kashy, and Bolger (1998), we conducted a formal test of mediation and found that the effect of mastery goals on catch-2 was significantly mediated through catch-1 ( $z = 6.65, p < .01$ ). Figure 3 depicts this mediation analysis.

**Hold.** The basic short-term model was significant,  $F(15, 842) = 44.31, p < .001$  ( $R^2 = .44$ ). There was a significant main effect of initial interest,  $F(1, 842) = 164.61, p < .001$  ( $\beta = .42$ ), indicating that students who entered the class with a high level of interest reported higher levels of hold. A significant main effect of

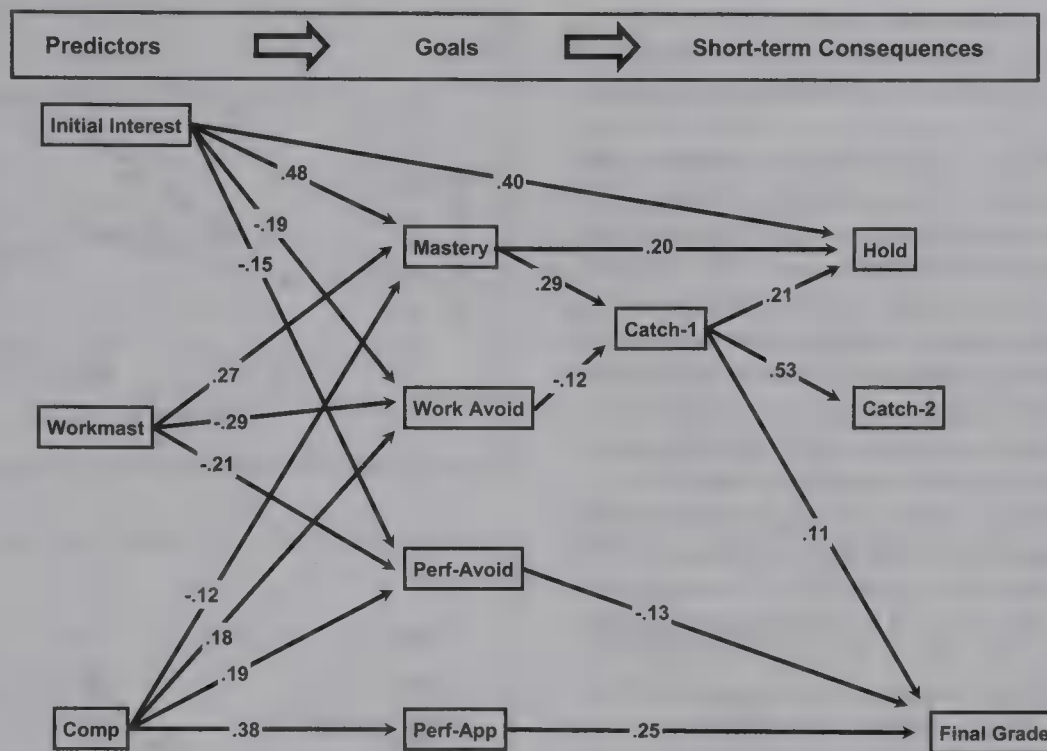


Figure 2. Predictors of short-term consequences. All paths represented are significant ( $p < .01$ ). Path coefficients are standardized regression coefficients. For clarity of presentation, this figure does not show the effects of gender on goal adoption, because gender had no additional effects over time. Instructor effects are not shown in this model, but both gender and instructor effects were controlled for in the model depicted. Workmastery = workmastery; Comp = competitiveness; Mastery = mastery goals; Perf-Avoid = performance-avoidance goals; Perf-App = performance-approach goals; Work Avoid = work-avoidance goals.

mastery goals,  $F(1, 842) = 57.08, p < .001$  ( $\beta = .26$ ), indicated that individuals who had adopted mastery goals early in the semester also reported that they found psychology meaningful and interesting at the end of the semester.

The catch-1 short-term model accounted for significantly more variance than the basic short-term model ( $R^2_{\text{change}} = .03, p < .001$ ), revealing that the effect of catch-1 was significant,  $F(1, 841) = 48.56, p < .001$  ( $\beta = .21$ ). Students who reported catch at Time 2 were more likely to report hold at Time 3. The effects of initial interest and mastery goals remained significant ( $\beta$ s = .40 and .20, respectively). The direct effect of mastery goals was reduced in magnitude, and a formal test of mediation indicated that the effect of mastery goals was significantly mediated through catch-1 ( $z =$

5.01,  $p < .01$ ; see Figure 3). Considered together, these results suggest that catch, initial interest, and mastery goals were all unique contributors to students' interest (hold) at the end of the course (see Figure 2).

**Final grade.** The basic short-term model was significant,  $F(15, 842) = 5.70, p < .001$  ( $R^2 = .09$ ); and there were significant main effects of performance-approach goals,  $F(1, 842) = 45.61, p < .001$  ( $\beta = .25$ ); and performance-avoidance goals,  $F(1, 842) = 14.33, p < .001$  ( $\beta = -.14$ ). Students who adopted performance-approach goals achieved higher grades in their introductory psychology course, and students who adopted performance-avoidance goals earned lower grades.

The catch-1 short-term model accounted for significantly more variance than the basic model ( $R^2_{\text{change}} = .01, p < .01$ ). This was due to the significant effect of catch-1,  $F(1, 841) = 8.39, p < .01$  ( $\beta = .11$ ), indicating that, as predicted, students who reported higher levels of catch-1 attained higher grades in the class. The effects of performance-approach and performance-avoidance goals remained significant ( $\beta$ s = .25 and -.13, respectively). These results are depicted in Figure 2.

### Predictors of Long-Term Outcomes

**Regression models for long-term analyses.** In this series of multiple regression analyses, we tested long-term effects in two steps. First, to investigate the direct effects of initial interest, achievement orientation, and goals on the long-term outcomes, we used the 15-term basic model established in the short-term analyses to predict psychology courses taken, major, and psychology GPA. However, because of the great variability in the total number

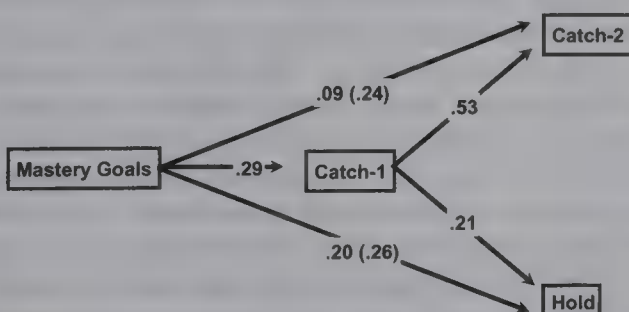


Figure 3. Mediation of catch-2 and hold by catch-1. Schematic diagram showing mediated effects, with direct paths from the original basic short-term model shown in parentheses. The path from mastery goals to catch-2 was no longer significant in the catch-1 short-term model, but the path from mastery goals to hold remained significant.



of academic credits taken over this time period ( $M = 79.41$ ,  $SD = 18.04$ ), we also controlled for total number of credits taken in all analyses, resulting in a 16-term direct effects long-term model (see Table 2).

Second, to examine the long-term effects in the context of the short-term effects documented earlier, we added the short-term outcome measures to the direct effects long-term model to test for mediated or indirect effects. Specifically, we added catch-1, catch-2, hold, and final grade to the direct effects model. In addition, we tested the six two-way interactions between catch-1, catch-2, hold, and final grade, but preliminary testing revealed that only the Hold  $\times$  Grade interaction was significant on any measure, and thus the other interactions were trimmed from the model, resulting in a 21-term final long-term model (see Table 2).

**Direct effects on psychology courses taken.** The direct effects long-term model was significant,  $F(16, 841) = 9.58$ ,  $p < .001$  ( $R^2 = .15$ ). There was a significant main effect of initial interest,  $F(1, 841) = 27.82$ ,  $p < .001$  ( $\beta = .21$ ); a significant main effect of background knowledge,  $F(1, 841) = 12.21$ ,  $p < .01$  ( $\beta = .12$ ); and a significant interaction between initial interest and background knowledge,  $F(1, 841) = 12.46$ ,  $p < .001$  ( $\beta = .11$ ). Students who entered the class with higher levels of initial interest were more likely to take more psychology classes if they also had more background in psychology before they took the course ( $\hat{Y} = 10.18$ ), compared to students who did not have much background ( $\hat{Y} = 5.37$ ), and relative to students low in initial interest with high ( $\hat{Y} = 3.41$ ) or low ( $\hat{Y} = 3.48$ ) background knowledge. This interaction is graphed in Figure 4A.

**Mediated and indirect effects on psychology courses taken.** The final long-term model, which included the short-term outcomes, was significant,  $F(21, 836) = 9.02$ ,  $p < .001$  ( $R^2 = .19$ ), and accounted for significantly more variance than the direct effects long-term model,  $F(5, 836) = 6.24$ ,  $p < .001$ . In this model, the main effect of hold was significant,  $F(1, 836) = 14.92$ ,  $p < .001$  ( $\beta = .20$ ). However, the effect of initial interest was reduced in size ( $\beta = .13$ , from .21 in the direct effects model), and a formal test of mediation revealed that the initial interest effect was partially mediated through hold ( $z = 3.67$ ,  $p < .01$ ). The interaction between initial interest and background knowledge remained significant ( $\beta = .11$ ). Finally, the interaction between hold and final grade was significant,  $F(1, 836) = 11.68$ ,  $p < .01$  ( $\beta = .11$ ), indicating that students took more psychology classes when they were high in hold and had high grades in their introductory course ( $\hat{Y} = 9.08$ ) relative to students high in hold with low grades ( $\hat{Y} = 6.27$ ) and relative to students low in hold with either high ( $\hat{Y} = 2.87$ ) or low ( $\hat{Y} = 4.39$ ) grades. This interaction is graphed in Figure 5A. Figure 6 depicts the paths to the three long-term variables (courses taken, psychology major, and psychology GPA) estimated from the final long-term model.

**Direct effects on psychology major.** For these analyses, our sample was limited to the 770 students who had declared majors during their academic careers. Although logistic regression is the most appropriate analysis for a dichotomous measure such as majoring (or not) in psychology, Rosenthal and Rosnow (1991) noted that multiple regression procedures yield accurate results for dichotomous dependent variables as long as the sample size is large and the split between the occurrences of each dichotomous outcome is not extreme (as was the case here). For consistency with the rest of the Results section, we report multiple regression

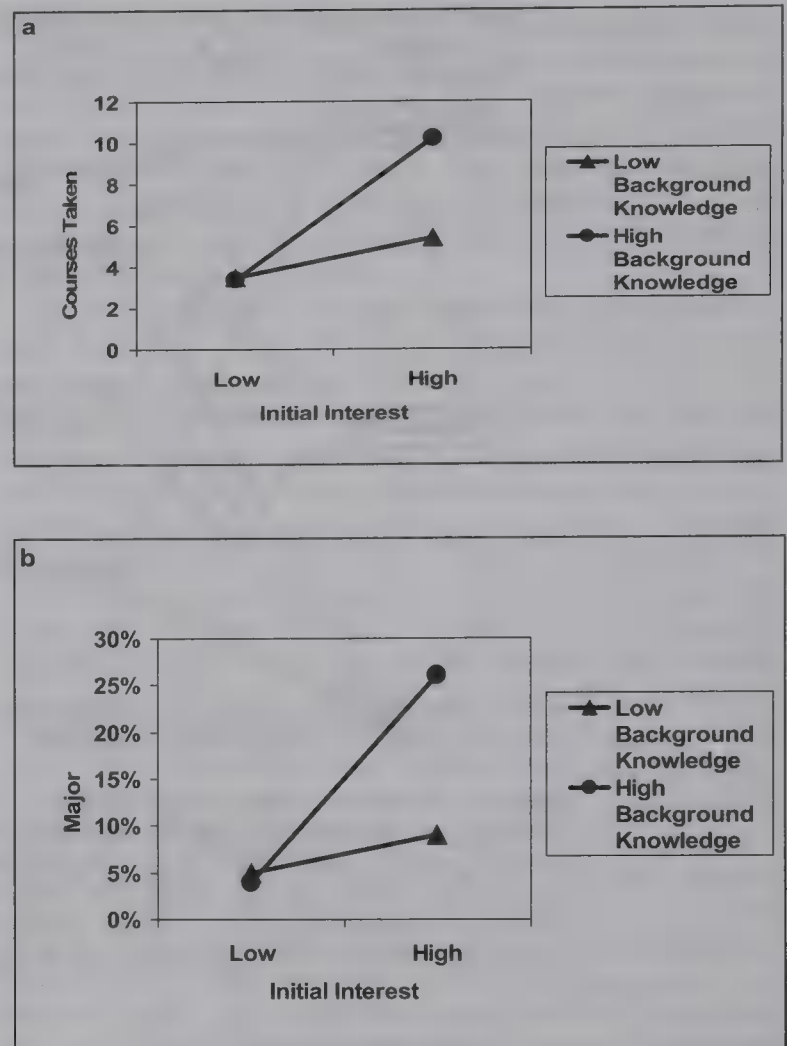


Figure 4. Predicted values for Initial Interest  $\times$  Background Knowledge interactions for courses taken (A) and for majoring in psychology (B). We computed predicted values for representative high and low groups (1 SD above and below the mean) for initial interest and background knowledge, respectively, from the regression equations using the unstandardized regression coefficients.

analyses, but we also conducted a series of logistic regressions and obtained the same pattern of results.

We first regressed psychology major on the direct effects long-term model. This model was significant,  $F(16, 753) = 6.76$ ,  $p < .001$  ( $R^2 = .13$ ). There were significant main effects of initial interest,  $F(1, 753) = 21.42$ ,  $p < .001$  ( $\beta = .20$ ); and background knowledge,  $F(1, 753) = 12.04$ ,  $p < .01$  ( $\beta = .12$ ). In addition, the interaction of initial interest and background knowledge was significant,  $F(1, 753) = 13.79$ ,  $p < .001$  ( $\beta = .13$ ). Students who entered the class with higher levels of initial interest were more likely to major in psychology if they also had more background in psychology before they took the course ( $\hat{Y} = 0.26$ ), compared to students who did not have much background ( $\hat{Y} = 0.09$ ) and relative to students low in initial interest with high ( $\hat{Y} = 0.04$ ) or low ( $\hat{Y} = 0.05$ ) background. This interaction is graphed in Figure 4B.

**Mediated and indirect effects on psychology major.** The final long-term model was significant,  $F(21, 748) = 6.19$ ,  $p < .01$  ( $R^2 = .15$ ), and it accounted for significantly more variance than the direct effects model,  $F(5, 748) = 3.94$ ,  $p < .01$ . In this model, the main effect of hold was significant,  $F(1, 748) = 11.89$ ,  $p =$

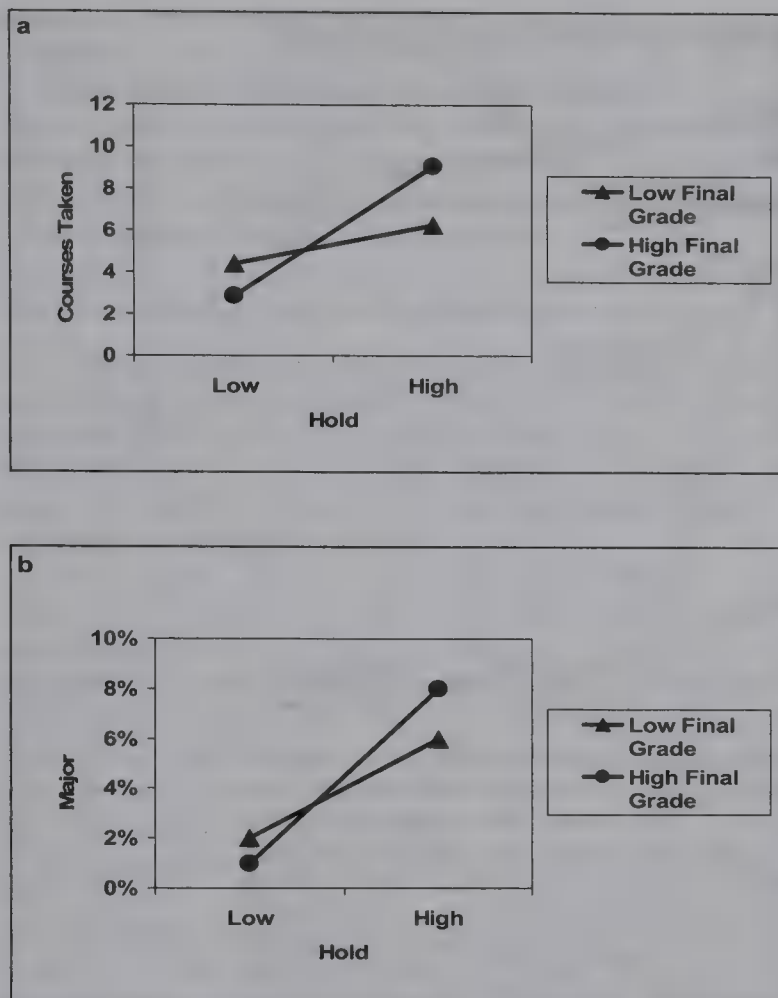


Figure 5. Predicted values for Hold  $\times$  Final Grade interactions for courses taken (A) and for majoring in psychology (B). We computed predicted values for representative high and low groups (1 SD above and below the mean) for hold and final grade, respectively, from the regression equations using the unstandardized regression coefficients.

.001 ( $\beta = .19$ ), indicating that students who reported higher levels of hold during the introductory class were more likely to eventually major in psychology. The initial interest effect remained significant but was reduced in size ( $\beta = .12$ , from .20 in the direct effects model). A formal test of mediation revealed that this drop in beta was significant, suggesting partial mediation through hold ( $z = 3.34$ ,  $p < .01$ ). The effect of background and the interaction of initial interest with background remained significant ( $\beta$ s = .11 and .12, respectively). The interaction between hold and final grade was not significant,  $F(1, 749) = 4.49$ ,  $p < .05$  ( $\beta = .07$ ), but was in the same direction as the significant interaction reported for psychology courses taken; this is depicted in Figure 5B. Figure 6 shows the significant paths to psychology major estimated from the final long-term model.

**Direct effects on psychology grades.** For the psychology GPA analysis, our sample was limited to the 341 students who actually enrolled in additional psychology classes. The direct effects long-term model was significant,  $F(16, 324) = 2.56$ ,  $p < .01$  ( $R^2 = .11$ ), and the main effect of performance-approach goals was significant,  $F(1, 324) = 12.84$ ,  $p < .001$  ( $\beta = .21$ ), indicating that students who adopted performance-approach goals obtained higher grades in subsequent psychology classes. In addition, the main effect of courses taken was significant,  $F(1, 324) = 6.67$ ,  $p < .01$

( $\beta = .14$ ), suggesting that students who took more credits overall performed better in their psychology coursework.

**Mediated and indirect effects on psychology grades.** The final long-term model was significant,  $F(21, 319) = 12.47$ ,  $p < .001$  ( $R^2 = .45$ ), and accounted for significantly more variance than the direct effects long-term model,  $F(5, 319) = 39.36$ ,  $p < .001$ . The main effect of final grade was significant,  $F(1, 319) = 191.52$ ,  $p < .001$  ( $\beta = .62$ ). Students who received higher grades in Introductory Psychology received higher grades in their later psychology courses. The main effect of performance-approach goals was no longer significant,  $F(1, 319) = 2.38$ ,  $p = .12$  ( $\beta = .07$ ), suggesting that the direct performance-approach goal effect on psychology grades was partially mediated by grades in the introductory course. Again, this was confirmed with a formal test of mediation ( $z = 6.08$ ,  $p < .01$ ). Figure 6 shows the significant paths to psychology GPA estimated from the final long-term model.

### Reciprocal Effects of Interest and Performance

The ordering of Time 2 (catch-1) and Time 3 (catch-2 and hold) interest measures in relation to students' exams and receipt of performance feedback allowed us to examine the reciprocal relations among interest and performance. In the analyses reported previously, we found that catch-1, measured before any exams in the course, was a significant predictor of final grade, suggesting that early situational interest predicted performance. In the analyses reported in the following sections, we also tested whether catch-2 or hold predicted final grade to determine whether situational interest developed later in the semester predicted performance. Finally, to examine the reciprocal effects of performance and interest, we tested the effects of catch-1 on initial exam performance and then the effects of initial exam performance on catch-2 and hold.

**Effects of catch-2 and hold on final grade.** We tested a regression model on final grade that included the catch-1 short-term model with the addition of catch-2 and hold. In other words, we tested the incremental effects of catch-2 and hold, over catch-1, on final grade. This model,  $F(18, 839) = 7.64$ ,  $p < .001$  ( $R^2 = .14$ ), accounted for significantly more variance than the catch-1 short-term model ( $p < .001$ ). The effects of catch-2,  $F(1, 839) = 12.89$ ,  $p < .001$  ( $\beta = .17$ ); and hold,  $F(1, 839) = 6.66$ ,  $p < .01$  ( $\beta = .13$ ), were both significant, suggesting that catch-2 and hold both predicted graded performance in the class. The effect of performance-approach goals remained significant ( $\beta = .24$ ), but the effect of catch-1 was reduced from .12 to  $-.01$ . Formal tests of mediation revealed that the effect of catch-1 on final grade was mediated through both catch-2 ( $z = 3.54$ ,  $p < .01$ ) and hold ( $z = 2.39$ ,  $p < .05$ ).

Although these results suggest that catch-2 and hold promoted performance in the course, these analyses were confounded by the fact that students were aware of their performance on two exams at the point in time that we measured catch-2 and hold. It is therefore possible that performance on early exams actually influenced self-reports of catch-2 and hold. This potential confound does not apply to catch-1, however, because we measured catch-1 before any exams or feedback in the course. The first exam was taken after the assessment of catch-1 and before the assessment of catch-2 and hold, allowing us to assess the effects of first exam performance on interest, controlling for initial interest and catch-1.



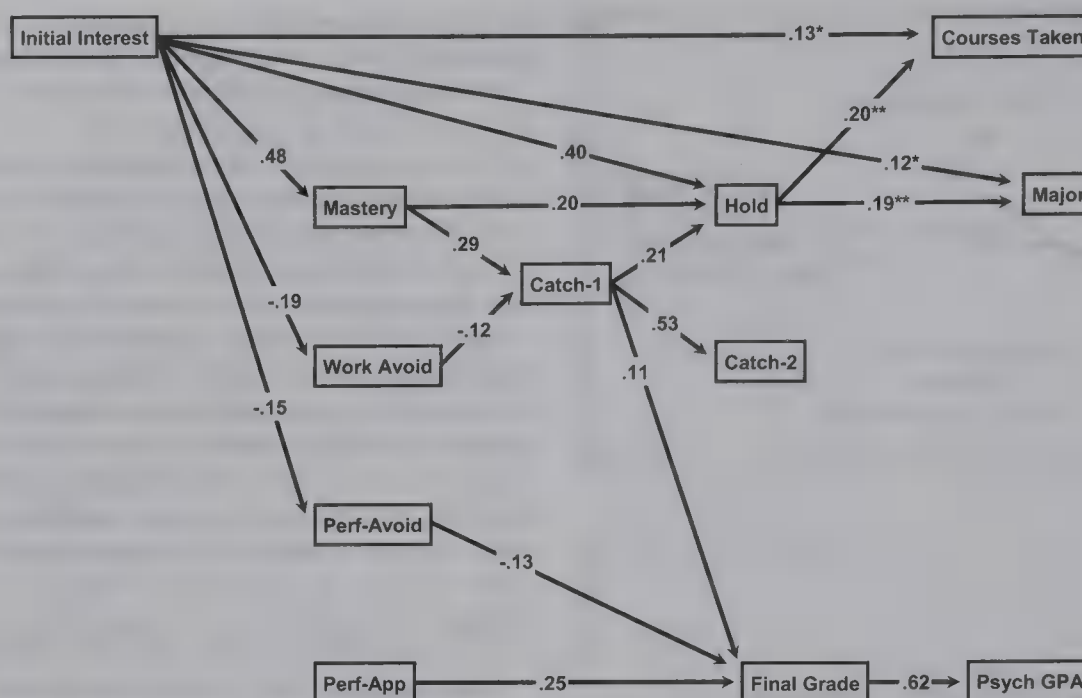


Figure 6. Predictors of long-term consequences. All paths represented are significant ( $p < .01$ ). Path coefficients are standardized regression coefficients. Paths with asterisks indicate significant interactions. Paths with a single asterisk reflect significant interaction effects between initial interest and background knowledge, whereas paths with two asterisks indicate significant interactions between hold and final grade. Separate coefficients were derived from the overall regression equation for individuals 1 *SD* above or below the mean (Aiken & West, 1991; Judd & McClelland, 1989). Specifically, the significant effect of initial interest on courses taken ( $\beta = .13$ ) and psychology major ( $\beta = .12$ ) was qualified by a significant interaction with background knowledge, such that the effect of initial interest on courses taken and psychology major was stronger for students with higher levels of background knowledge ( $\beta$ s = .24 and .33, respectively) than for students with less background knowledge ( $\beta$ s = .02 and .07, respectively). The significant effect of hold on courses taken ( $\beta = .20$ ) and major ( $\beta = .19$ ) was qualified by a significant interaction with final grade, such that the effect of hold on courses taken and psychology major was stronger for students with higher grades in Introductory Psychology ( $\beta$ s = .31 and .26, respectively) than it was for students with lower grades ( $\beta$ s = .09 and .12, respectively). For clarity of presentation, significant gender and instructor effects are not depicted, but these variables were controlled in these analyses. Mastery = mastery goals; Perf-Avoid = performance-avoidance goals; Perf-App = performance-approach goals; Work Avoid = work-avoidance goals.

We therefore examined the reciprocal effects of interest and performance by examining the predictors and correlates of performance on the first exam.

*Effects of catch-1 on early performance.* In all classes, catch-1 was measured at least 1 week prior to the first exam, allowing us to test catch-1 as an uncontaminated predictor of exam performance. When we regressed first exam score on the catch-1 short-term model, the overall model was significant,  $F(16, 841) = 4.61$ ,  $p < .001$  ( $R^2 = .08$ ). There were significant effects of performance-approach goals,  $F(1, 841) = 44.75$ ,  $p < .001$  ( $\beta = .25$ ); performance-avoidance goals,  $F(1, 841) = 8.62$ ,  $p < .01$ , ( $\beta = -.11$ ); and catch-1,  $F(1, 841) = 9.59$ ,  $p < .01$  ( $\beta = .12$ ). Students who adopted performance-approach goals, as well as students who enjoyed the course lectures, performed better on the first exam; students who adopted performance-avoidance goals performed worse.

*Effects of early performance on catch-2 and hold.* We then tested first exam performance as a predictor of catch-2 and hold, which were measured at least 2 months after the first exam. We added first exam score to the catch-1 short-term model to predict catch-2 and hold.

On catch-2, the overall model was significant,  $F(17, 840) = 38.49$ ,  $p < .001$  ( $R^2 = .44$ ), and accounted for significantly more variance than the catch-1 short-term model reported earlier ( $R^2_{\text{change}} = .02$ ,  $p < .001$ ). The effect of catch-1 was significant,  $F(1, 840) = 283.55$ ,  $p < .001$  ( $\beta = .51$ ), suggesting that students who reported higher levels of catch-1 at Time 2 also reported higher levels of catch-2 at Time 3. The effect of the first exam score was significant,  $F(1, 840) = 31.11$ ,  $p < .001$  ( $\beta = .15$ ), suggesting that students who performed well on the first exam reported higher levels of catch-2 at Time 3, controlling for the effects of catch-1.

On hold, the overall model including first exam score was significant,  $F(17, 840) = 47.59$ ,  $p < .001$  ( $R^2 = .49$ ), and accounted for significantly more variance than the catch-1 short-term model reported earlier ( $R^2_{\text{change}} = .02$ ,  $p < .001$ ). The main effects of initial interest,  $F(1, 840) = 164.17$ ,  $p < .001$  ( $\beta = .40$ ); mastery goals,  $F(1, 840) = 36.64$ ,  $p < .001$  ( $\beta = .21$ ); and catch-1,  $F(1, 840) = 41.72$ ,  $p < .001$  ( $\beta = .19$ ), were significant. In addition, the effect of first exam performance was significant,  $F(1, 840) = 31.21$ ,  $p < .001$  ( $\beta = .14$ ), suggesting that students who performed well on the first exam reported higher levels of hold at Time 3. In

sum, these results indicate that performance on an early exam can contribute to catch and hold independent of the other interest processes already documented. Considered together with the effects of catch-1 on early and later performance, these findings illustrate the dynamic relationship between interest and performance over time. Figure 7 presents a path model of the predictors of early exam performance, catch-2, and hold.

#### Ancillary Analyses on General Performance Measures

As we have done in our prior research, we also examined the predictors of general academic performance in both the short (semester GPA) and the long term (subsequent GPA). To test the direct effects of individual interest, achievement orientation, and goals, we used the 15-term basic short-term model to predict semester GPA and subsequent GPA.

**Semester GPA.** The basic short-term model was significant,  $F(15, 842) = 4.22, p < .001 (R^2 = .07)$ . Main effects of performance-approach goals,  $F(1, 842) = 36.00, p < .001 (\beta = .22)$ ; and performance-avoidance goals,  $F(1, 842) = 7.83, p < .01 (\beta = -.10)$ , showed that students who adopted performance-approach goals in their psychology class attained higher grades in all of their classes that semester, whereas students who adopted performance-avoidance goals earned lower grades.

**Subsequent GPA.** The basic short-term model was significant,  $F(15, 842) = 4.55, p < .001 (R^2 = .08)$ . There were main effects of performance-approach goals,  $F(1, 842) = 19.57, p < .001 (\beta = .16)$ ; performance-avoidance goals,  $F(1, 842) = 7.47, p < .01 (\beta = -.10)$ ; and work-avoidance goals,  $F(1, 842) = 7.59, p < .01 (\beta = -.11)$ . Students who adopted performance-approach goals in Introductory Psychology had higher grades over the course of their

academic careers, whereas students who adopted performance-avoidance or work-avoidance goals had lower grades.

#### Discussion

By situating our analyses of interest development within the context of achievement goals research, we gained insight into the role of initial interest and mastery goals in promoting the development of situational and continuing interest in an academic domain. Our results clearly indicate that interest and mastery goals are reciprocally related over time. Initial interest, measured at the outset of a college course, predicted mastery goal adoption, situational interest during the course, as well as continued interest measured several semesters later. Moreover, the effects of initial interest on continued interest were partially mediated through mastery goals, suggesting that interest can deepen over time because initial interest motivates individuals to take a task-focused approach to the material, with a desire to learn more about the topic (Renninger, 2000). This mastery goal approach then facilitates the further development and deepening of interest as students become engaged with the course material and learn more about the topic. Thus, mastery goals can be viewed as both a product and predictor of interest, or as a mediating mechanism for the continued development of interest in a topic.

We had originally contrasted these two perspectives—that interest promotes the adoption of mastery goals, and that mastery goals promote the development of interest—as separate alternatives, but it is clear from our longitudinal results that both perspectives have validity. When individuals enter a situation with interest in the topic, they may be motivated to learn more about it (i.e., adopt a mastery goal), and they may also develop more

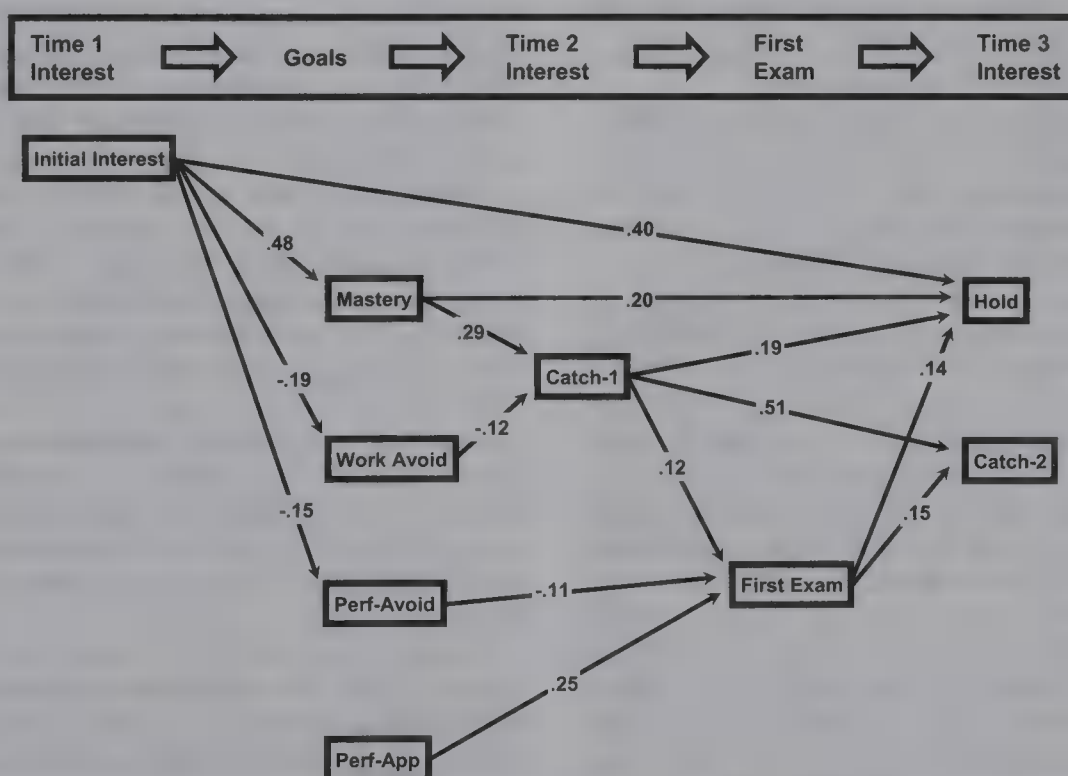


Figure 7. Reciprocal effects of interest and performance. All paths represented are significant ( $p < .01$ ). Path coefficients are standardized regression coefficients. Mastery = mastery goals; Perf-Avoid = performance-avoidance goals; Perf-App = performance-approach goals; Work Avoid = work-avoidance goals.



interest when they approach a task with a mastery goal. These two perspectives can be reconciled by distinguishing the individual interest that students bring to tasks (which can promote mastery goal adoption) from the situational interest that may develop as a result of task engagement (and mastery goal pursuit). Our findings suggest that interest must be conceptualized as an ongoing process and studied over time to elucidate the processes through which initial interest affects goal adoption and continued interest, as well as the processes through which goal adoption influences the development and deepening of interest.

Our results offer strong support for the four-phase model of interest development recently advanced by Hidi and Renninger (2006), which identifies four phases of interest development and outlines the progression from situational interest to individual interest. According to their model, the first phase of interest development is a *triggered situational interest*, in which attention, liking, or involvement is initiated by an external cue. Our lecture-specific measures of catch correspond to this early phase of interest development. If sustained over time or contexts, triggered situational interest evolves into the second phase, a *maintained situational interest*, and this phase of interest may be reflected in our measures of hold, which assessed interest in the course material. A third phase, characterized as an *emerging individual interest*, may develop out of the second phase if individuals begin to value the object or topic beyond the situation that first stimulated their interest. In this phase, individuals may be predisposed to seek out opportunities on their own to reengage with the topic of interest. This phase of interest may be reflected in our behavioral measure of continued interest, which examined students' course choices after completion of the introductory course. The third phase of interest development can then lead to the fourth, a *well-developed individual interest*, which is associated with personal meaning, value, and knowledge (Renninger, 1990). Our psychology major variable, which represents a student's extensive experience with and investment in a domain of study, seems to correspond to a well-developed individual interest.

According to Hidi and Renninger (2006), each phase is characterized by varying amounts of affect, knowledge, and value, with a greater emphasis on stimulation, affect, and liking in the early phases and a greater emphasis on meaning, value, and stored knowledge in the later phases (Renninger, 1989, 1990, 1998). The progression from catch to hold to continued interest seen in our path models maps onto the deepening of interest as described by Hidi and Renninger. In particular, the fact that the effects of mastery goals on situational interest (catch-2 and hold) were mediated by early situational interest (catch-1) suggests the mechanism through which goals can promote situational interest. Adopting a mastery goal may lead individuals to actively engage in the learning experience so that they discover the interesting and stimulating aspects of the topic (catch-1, catch-2). In the college classroom, students with a mastery goal may be more likely to notice interesting details in a text or appreciate the stimulating qualities of a lecture presentation. Over time, this heightened attention may help students gain a greater appreciation of the topic more generally or see connections to their life. In other words, triggered situational interest may develop into a maintained situational interest if students discover the value or personal relevance of the material, as suggested by the significant effect of catch-1 on hold. Our analysis is consistent with the idea that interest must be

caught before it can be held (Dewey, 1913; Hidi & Baird, 1986; Mitchell, 1993).

We also found that the effects of initial interest and mastery goals on continued interest were mediated through hold, which supports the Hidi and Renninger (2006) model by documenting how interest can deepen over time when individuals perceive value and personal relevance in a task. It is important to note that neither catch-1 nor catch-2 predicted continued interest, replicating the results of Harackiewicz et al. (2000). Although catch was important in the early stages of interest development, hold proved to be more relevant to the deepening of interest, or the transformation of situational interest into an emerging individual interest (Mitchell, 1993; Renninger, 1992). The significant interactions of hold with final grade on courses taken and major further suggest that hold was a particularly strong predictor of interest for students who received higher grades in the course. These interactions replicate results reported by Harackiewicz, Barron, Tauer, and Elliot (2002) and suggest that continued interest in a domain depends on both interest and performance. Moreover, these effects represent indirect effects of performance-approach goals on courses taken and major, because grades were predicted by performance-approach goals. Doing well in the course could contribute to the stored value of the domain (Renninger, 1992) and might therefore facilitate reentry into situations that offer the opportunity to learn more about the topic (e.g., more psychology courses). Thus, both mastery and performance-approach goals are important in predicting continued interest in an academic domain, offering further support for the multiple goals model (Harackiewicz, Barron, Pintrich, et al., 2002).

The four-phase model also affords an analysis of the effects of initial interest on subsequent interest development. We might consider students who were high in initial interest in our study to have begun the course with either an emerging or a well-developed individual interest, depending on their level of background knowledge. The fact that initial interest and background knowledge were only moderately correlated ( $r = .21$ ) suggests that there was considerable variability in the degree of background knowledge associated with high initial interest. For some students, high levels of initial interest were paired with high levels of background knowledge, indicating a well-developed interest, whereas other students reported equally high interest but lower levels of background knowledge, suggesting that they were attracted to the topic but did not have much substantive experience with it (Renninger, 2000). This suggests that individuals entered this class at many different levels along the continuum of developing interest. The interactive effects of initial interest and background knowledge on courses taken and major indicate that initial interest was a particularly strong predictor of continued interest when paired with high levels of background knowledge, supporting Renninger's (2000) assertion that strong individual interest involves knowledge of the topic.

Among students who began the class without strong interest, however, the development of interest had to begin with situational interest. Because situational interest is externally supported, these individuals could only develop interest if features of the class or aspects of the course materials triggered and then maintained their situational interest. Consequently, the students who entered the course with low initial interest but who experienced triggered situational interest may have progressed to maintained situational



interest, and possibly even to emerging individual interest. Recent experimental research supports this analysis: Instructional enhancements designed to catch interest were found to be especially advantageous for those who entered a learning situation with low initial interest (Durik & Harackiewicz, 2007). In contrast, those students who entered the class with high initial interest may have bypassed triggered situational interest and progressed from a maintained situational interest to an emerging individual interest to the extent that they found value or personal meaning in the course material. Durik and Harackiewicz also found that instructional enhancements designed to hold interest were particularly effective for individuals high in initial interest, offering further experimental support for this analysis.

We also documented the processes through which interest predicted academic performance and showed that early situational interest was a predictor of students' final grades in the course, as well as their performance on the first exam in the course (Hidi & Baird, 1986). These findings are particularly significant because our early measure of situational interest (catch-1) was collected before any exams or feedback and was therefore uncontaminated by actual or perceived performance. Although these effects were relatively small ( $\beta$ s = .11 and .12), they do suggest that motivational variables can play an important role in predicting performance. In particular, these effects are indicative of indirect mastery goal effects, because although mastery goals did not predict performance, they did predict catch-1, which then predicted performance. Although not as large as the direct performance-approach goal effect on first exam performance and final grade, these effects suggest that mastery goals may indirectly promote performance in college courses through the interest processes they engender. Moreover, the analyses with first exam performance reveal the reciprocal effects of interest and performance, notably that early interest predicted exam performance and grades, and that early exam performance predicted subsequent situational interest (both catch and hold), controlling for initial interest. These results are consistent with the idea that students perform well on tasks they find interesting and that they become more interested in activities when they have performed well on them. When students perform well on exams, their sense of competence or self-efficacy may lead them to value and enjoy that activity (Wigfield & Eccles, 1992; Zimmerman, 1985). Again, two alternative perspectives (that interest influences performance, and that performance influences interest) can be resolved by studying these processes over time to elucidate the reciprocal relations among these important variables.

The effects of early situational interest on performance and subsequent situational interest (hold) indicate the importance of externally supported interest and suggest that educators may be able to influence performance and continued interest through situational interest. If triggered situational interest was the critical factor for engaging students who were initially uninterested in the material, then it is important to identify what types of classroom environments best promote triggered situational interest. It is not known whether instructional technology, the organization of the material, or the instructor's style triggered situational interest in this study, but understanding learners' responses to such situational variables will be critical for designing instructional contexts that maximize student interest and performance. Experimental research has begun to examine the instructional factors that affect

situational interest, such as whether the task is visually appealing and whether the utility value of the material is emphasized (Durik & Harackiewicz, 2007), but more work is needed to better understand how situational factors can be used in classroom environments to promote optimal motivation.

Moreover, future research should examine how situational and individual interests combine across longer spans of time and during different periods of personal development. Parents and peers may affect situational interest in activities over time, influencing the extent to which individuals have contact with various domains, adopt achievement goals, and identify them as meaningful and interesting. The influence of parents might be especially prominent for young children, whereas the influence of peers might be more prominent during adolescence. Although the present study begins to elucidate how interest and goals unfold over time within an introductory college course, it will be important to extend these analyses beyond the college classroom in future research.

## References

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- Archer, J. (1994). Achievement goals as a measure of motivation in university students. *Contemporary Educational Psychology*, 19, 430–446.
- Barron, K. E., & Harackiewicz, J. M. (2000). Achievement goals and optimal motivation: A multiple goals approach. In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 229–254). New York: Academic Press.
- Barron, K. E., & Harackiewicz, J. M. (2001). Achievement goals and optimal motivation: Testing multiple goal models. *Journal of Personality and Social Psychology*, 80, 706–722.
- Barron, K. E., & Harackiewicz, J. M. (2003). Revisiting the benefits of performance-approach goals in the college classroom: Exploring the role of goals in advanced college courses. *International Journal of Educational Research*, 39, 357–374.
- Bouffard, T., Boisvert, J., Vezeau, C., & Larouche, C. (1995). The impact of goal orientation on self-regulation and performance among college students. *British Journal of Educational Psychology*, 65, 317–329.
- Bouffard, T., Vezeau, C., & Bordeleau, L. (1998). A developmental study of the relation between combined learning and performance goals and students' self-regulated learning. *British Journal of Educational Psychology*, 68, 309–319.
- Brophy, J. E. (1983). Conceptualizing student motivation. *Educational Psychologist*, 18, 200–215.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Dewey, J. (1913). *Interest and effort in education*. Boston: Riverside Press.
- Durik, A. M., & Harackiewicz, J. M. (2007). Different strokes for different folks: Individual interest as a moderator of the effects of situational factors on task interest. *Journal of Educational Psychology*, 99, 597–610.
- Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, 95, 256–273.
- Elliot, A. J., & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology*, 72, 218–232.
- Harackiewicz, J. M., Barron, K. E., Carter, S. M., Lehto, A. T., & Elliot, A. J. (1997). Predictors and consequences of achievement goals in the college classroom: Maintaining interest and making the grade. *Journal of Personality and Social Psychology*, 73, 1284–1295.



- Harackiewicz, J. M., Barron, K. E., & Elliot, A. J. (1998). Rethinking achievement goals: When are they adaptive for college students and why? *Educational Psychologist*, 33, 1–21.
- Harackiewicz, J. M., Barron, K. E., Pintrich, P. R., Elliot, A. J., & Thrash, T. M. (2002). Revision of achievement goal theory: Necessary and illuminating. *Journal of Educational Psychology*, 94, 638–645.
- Harackiewicz, J. M., Barron, K. E., Tauer, J. M., Carter, S. M., & Elliot, A. J. (2000). Short-term and long-term consequences of achievement goals: Predicting interest and performance over time. *Journal of Educational Psychology*, 92, 316–330.
- Harackiewicz, J. M., Barron, K. E., Tauer, J. M., & Elliot, A. J. (2002). Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of Educational Psychology*, 94, 562–575.
- Helmreich, R. L., & Spence, J. T. (1978). The Work and Family Orientation Questionnaire: An objective instrument to assess components of achievement motivation and attitudes toward family and career. *JSAS Catalog of Selected Documents in Psychology*, 8, 355.
- Hidi, S. (1990). Interest and its contribution as a mental resource for learning. *Review of Educational Research*, 60, 549–571.
- Hidi, S., & Baird, W. (1986). Interestingness—a neglected variable in discourse processing. *Cognitive Science*, 10, 179–194.
- Hidi, S., & Berndorff, D. (1998). Situational interest and learning. In L. Hoffman, A. Krapp, K. A. Renninger, & J. Baumert (Eds.), *Interest and learning: Proceedings of the Second Conference on Interest and Gender* (pp. 74–90). Kiel, Germany: IPN.
- Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research*, 70, 151–179.
- Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, 41, 111–127.
- Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model comparison approach*. San Diego, CA: Harcourt Brace Jovanovich.
- Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. Gilbert, S. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (pp. 233–265). New York: McGraw-Hill.
- Krapp, A. (2002). Structural and dynamic aspects of interest development: Theoretical considerations from an ontogenetic perspective. *Learning and Instruction*, 12, 383–409.
- Linnenbrink-Garcia, L., Durik, A. M., Conley, A. M., Barron, K. E., Tauer, J. M., Karabenick, S. A., et al. (2007). Situational Interest Survey (SIS): An instrument to assess the role of situational factors in interest development (submitted for publication).
- Maehr, M. L. (1976). Continuing motivation: An analysis of a seldom considered educational outcome. *Review of Educational Research*, 46, 443–462.
- Malone, T. W., & Lepper, M. R. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. In R. Snow & M. Farr (Eds.), *Aptitude, learning, and instruction* (Vol. 3, pp. 223–253). Hillsdale, NJ: Erlbaum.
- Middleton, M. J., & Midgley, C. (1997). Avoiding the demonstration of lack of ability: An underexplored aspect of goal theory. *Journal of Educational Psychology*, 89, 710–718.
- Midgley, C., Maehr, M. L., Hicks, L., Roeser, R., Urdan, T., Anderman, E., & Kaplan, A. (1996). *The patterns of adaptive learning survey (PALS)*. Ann Arbor: University of Michigan.
- Mitchell, M. (1993). Situational interest: Its multifaceted structure in the secondary school mathematics classroom. *Journal of Educational Psychology*, 85, 424–436.
- Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological Review*, 91, 328–346.
- Nicholls, J. G. (1989). *The competitive ethos and democratic education*. Cambridge, MA: Harvard University Press.
- Pintrich, P. R. (2000). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology*, 92, 544–555.
- Pintrich, P. R. (2002). Future challenges and directions for theory and research on personal epistemology. In B. Hofer & P. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 389–414). Mahwah, NJ: Erlbaum.
- Pintrich, P. R., & DeGroot, E. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology*, 82, 33–40.
- Pintrich, P. R., & Garcia, T. (1991). Student goal orientation and self-regulation in the college classroom. In M. L. Maehr & P. R. Pintrich (Eds.), *Advances in motivation and achievement* (Vol. 7, pp. 371–402). Greenwich, CT: JAI Press.
- Rathunde, K. (1993). The experience of interest: A theoretical and empirical look at its role in adolescent talent development. In P. Pintrich & M. Maehr (Eds.), *Advances in motivation and achievement* (Vol. 8, pp. 59–98). Greenwich, CT: JAI Press.
- Renninger, K. A. (1989). Individual patterns in children's play interests. In L. Winegar (Ed.), *Social interaction and the development of children's understanding* (pp. 147–172). Westport, CT: Ablex.
- Renninger, K. A. (1990). Children's play interests, representation, and activity. In R. Fivush & J. Hudson (Eds.), *Knowing and remembering in young children* (pp. 127–165). New York: Cambridge University Press.
- Renninger, K. A. (1992). Individual interest and development: Implications for theory and practice. In A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 361–395). Hillsdale, NJ: Erlbaum.
- Renninger, K. A. (1998). The role of interest and gender in learning: An overview of research on preschool and elementary-school-aged children. In L. Hoffman, A. Krapp, K. A. Renninger, & J. Baumert (Eds.), *Interest and learning: Proceedings of the Second Conference on Interest and Gender* (pp. 165–174). Kiel, Germany: IPN.
- Renninger, K. A. (2000). Individual interest and its implications for understanding intrinsic motivation. In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 373–404). New York: Academic Press.
- Rosenthal, R., & Rosnow, R. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.
- Schiefele, U. (1991). Interest, learning and motivation. *Educational Psychologist*, 26, 299–323.
- Spence, J. T., & Helmreich, R. L. (1983). Achievement-related motives and behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives: Psychological and sociological approaches* (pp. 7–74). San Francisco: Freeman.
- Spence, J. T., Pred, R. S., & Helmreich, R. L. (1989). Achievement strivings, scholastic aptitude, and academic performance: A follow-up to "Impatience versus achievement strivings in the Type A pattern." *Journal of Applied Psychology*, 74, 176–178.
- Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review*, 12, 265–310.
- Zimmerman, B. J. (1985). The development of "intrinsic" motivation: A social learning analysis. *Annals of Child Development*, 2, 117–160.

## Appendix

## Scale Items and Reliabilities

## Time 1 Items

Initial interest ( $\alpha = .90$ )

I've always been fascinated by psychology.

I chose to take Introductory Psychology because I'm really interested in the topic.

I'm really excited about taking this class.

I'm really looking forward to learning more about psychology.

I think the field of psychology is an important discipline.

I think what we will study in Introductory Psychology will be important for me to know.

I think what we will study in Introductory Psychology will be worthwhile to know.

Background knowledge ( $\alpha = .90$ )

This class is my first exposure to the field of psychology. (reversed)

I already have some background in psychology (e.g., I studied psychology in another class or did reading on my own).

I have very little experience with psychology. (reversed)

Workmastery (Helmreich & Spence, 1978; Spence & Helmreich, 1983;  $\alpha = .80$ )

I like to work hard.

There is satisfaction in a job well done.

I more often attempt tasks that I'm not sure I can do than tasks I believe I can do.

I would rather do something at which I feel confident and relaxed than something which is challenging and difficult.

I find satisfaction in working as hard as I can.

I would rather learn easy and fun games than difficult thought games.

Once I undertake a task I persist.

I find satisfaction in exceeding my previous performance even if I don't outperform others.

I like to be busy all the time.

Part of my enjoyment in doing things is improving my past performance.

When a group I belong to plans an activity, I would rather direct it myself than just help out and have someone else organize it.

I prefer to work in situations that require a high level of skill.

It is important to me to do my work as well as I can even if it isn't popular with my peers.

If I am not good at something, I would rather keep struggling to master it than move onto something I may be good at.

Competitiveness (Helmreich & Spence, 1978; Spence & Helmreich, 1983;  $\alpha = .75$ )

I feel that winning is important in both work and games.

I try harder when I'm in competition with other people.

It annoys me when other people perform better than I do.

I enjoy working in situations involving competition.

It is important to me to perform better than others on a task.

## Time 2 Items

Mastery goal orientation ( $\alpha = .87$ )

The most important thing for me in this course is to understand the content as thoroughly as possible.

Mastering the material in Introductory Psychology is important to me.

I want to learn as much as possible in this class.

I like it best when something I learn in this course makes me want to find out more.

In a class like this, I prefer course material that really challenges me so I can learn new things.

My goal in this class is to learn as much as I can about this topic.

In a class like this, I prefer course material that arouses my curiosity even if it is difficult to learn.

Performance-approach goal orientation ( $\alpha = .87$ )

It is important for me to do well compared to others in this class.

I don't care about how I do compared to the other students in this class. (reversed)

I want to do better than other students in this class.



My goal in this class is to get a better grade than most of the other students.

Performance-avoidance goal orientation ( $\alpha = .78$ )

I just want to avoid getting a low grade in this class.

I just want to avoid doing poorly in this class.

Work avoidance goal orientation ( $\alpha = .90$ )

I want to get through this course by doing the least amount of work possible.

I want to do as little work as possible in this class.

I don't want to work hard in this class. (reversed)

Catch-1 ( $\alpha = .85$ )

I look forward to coming to this class.

Lectures in this class really seem to drag on forever. (reversed)

I think the lectures are interesting.

The lectures in this class are entertaining.

Time 3 Items

Catch-2 ( $\alpha = .91$ )

I like my instructor.

The lectures in this class really seem to drag on forever. (reversed)

I don't like the lectures very much. (reversed)

I enjoy coming to lecture.

The lectures in this class aren't very interesting. (reversed)

Hold ( $\alpha = .95$ )

To be honest, I just don't find psychology interesting. (reversed)

I think the field of psychology is very interesting.

Psychology fascinates me.

I'm excited about psychology.

I think what we are learning in this course is important.

I think what we are studying in Introductory Psychology is useful for me to know.

I find the content of this course personally meaningful.

I see how I can apply what we are learning in Introductory Psychology to real life.

I think the field of psychology is an important discipline.

Received September 21, 2006

Revision received June 7, 2007

Accepted June 12, 2007 ■

# A Response to Recent Reanalyses of the National Reading Panel Report: Effects of Systematic Phonics Instruction Are Practically Significant

Karla K. Stuebing, Amy E. Barth, Paul T. Cirino, David J. Francis, and Jack M. Fletcher  
University of Houston

The authors examine the reassessments of the National Reading Panel (NRP) report (National Institute of Child Health and Human Development, 2000) by G. Camilli, S. Vargas, and M. Yurecko (2003); G. Camilli, P. M. Wolfe, and M. L. Smith (2006); and D. D. Hammill and H. L. Swanson (2006) that disagreed with the NRP on the magnitude of the effect of systematic phonics instruction. Using the coding of the NRP studies by Camilli et al. (2003, 2006), multilevel regression analyses show that their findings do not contradict the NRP findings of effect sizes in the small to moderate range favoring systematic phonics. Extending Camilli et al. (2003, 2006), the largest effects are associated with reading instruction enhanced with components that increase comprehensiveness and intensity. In contrast to Hammill and Swanson, binomial effect size displays show that effect sizes of the magnitude found for systematic phonics by the NRP are meaningful and could result in significant improvement for many students depending on the base rate of struggling readers and the size of the effect. Camilli et al. (2003, 2006) and Hammill and Swanson do not contradict the NRP report, concurring in supporting comprehensive approaches to reading instruction.

**Keywords:** reading instruction, phonics, National Reading Panel, meta-analysis

The report of the National Reading Panel (NRP; National Institute of Child Health and Human Development [NICHD], 2000), a congressionally mandated effort to synthesize research on effective instructional methods for teaching children to read, continues to generate controversy. The NRP report was completed by a committee under the direction of the NICHD in collaboration with the U.S. Department of Education. Despite the use of an empirical approach to the synthesis of research results (meta-analysis), with two peer-reviewed articles representing the syntheses of phonics and phonological awareness instruction (Ehri, Nunes, Stahl, & Willows, 2001; Ehri, Nunes, Willows, et al., 2001), the report, and especially the part involving phonics, has been highly scrutinized since it was released in 2000. Several reviewers have disagreed with NRP conclusions supporting the efficacy of systematic phonics instruction over other approaches to teaching phonics (Allington, 2002; Garan, 2001; see responses by Cooper, 2005; Shanahan, 2004).

Recently, two critiques of the phonics meta-analysis of the NRP report were published in a special issue of the *Elementary School*

*Journal* (Camilli, Wolfe, & Smith, 2006; Hammill & Swanson, 2006). Camilli et al. (2006) provided a second reanalysis of the studies coded by the NRP. In their first reanalysis of the NRP report, Camilli, Vargas, and Yurecko (2003) concluded that the actual effect size from studies involving systematic phonics instruction in the NRP report was  $d = 0.24$  when studies were weighted equally (i.e., without correction for sample size) and  $d = 0.188$  when studies were weighted by a combination of equal representation and sample size. These estimates were much lower than the  $d = 0.41$  reported by the NRP for end-of-training outcomes. In Camilli et al. (2006), more extensive coding of study characteristics was provided, along with appropriate multilevel analysis techniques, leading to  $d = 0.123$  for systematic phonics, which was characterized as not significant and a “weak intervention” (p. 31).

In a similar vein, Hammill and Swanson (2006) converted the estimates of effect size in the NRP report to metrics ( $R^2$ ) that represent the amount of explained variance. Although acknowledging that “94% of the  $d$ ’s supported the superiority of phonics instruction over other approaches,” they went on to observe that “Cohen would describe 65% of these significant  $d$ ’s as small” (p. 19). Converting the  $d$ s to  $r$ s yielded an overall  $r = .21$ , or  $R^2$  of .04, “suggesting that 96% of the variance in reading achievement can be attributed to factors other than the systematic phonics instruction” (p. 18). The authors concluded that “for all practical purposes, the advantages of phonics versus nonphonics instruction have not been demonstrated” (p. 25).

## Theoretical and Pedagogical Issues

Underlying the controversy over phonics and its role in reading instruction is a set of theoretical issues about learning to read that relate directly to how the alphabetic principle is taught (American

---

Karla K. Stuebing, Amy E. Barth, Paul T. Cirino, and David J. Francis, Department of Psychology; Texas Institute for Measurement, Evaluation and Statistics; and the Texas Center for Learning Disabilities, University of Houston; Jack M. Fletcher, Department of Psychology and the Texas Center for Learning Disabilities, University of Houston.

Grant P50 HD052117 from the National Institute of Child Health and Human Development to the Texas Center for Learning Disabilities supported this article. David J. Francis was a methodological consultant to the National Reading Panel (NRP). No other authors were involved with the NRP and its report.

Correspondence concerning this article should be addressed to Jack M. Fletcher, Department of Psychology, University of Houston TMC Annex, 2151 West Holcombe Boulevard, Suite 222, Houston, TX 77204-5355. E-mail: jackfletcher@uh.edu



Federation of Teachers, 1999; Pressley, 2005; Stanovich, 2000). Unlike in previous periods, the current discussion is rarely about whether any instruction involving the alphabetic principle should be provided, but (a) how systematically instruction should be conducted to ensure that all students have adequate knowledge of the alphabetic principle, and (b) the extent to which students are better served by opportunities to make inferences and develop their own understanding of the role of the alphabetic principle versus opportunities in which the alphabetic principle is directly taught by the teacher (Allington, 2002; Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001). These differences are often represented as a dichotomy comparing a scripted curriculum with a defined scope and sequence versus a curriculum that encourages discovery of the alphabetic principle through immersion in literature.

Underlying this dichotomy is a continuum that reflects the extent to which the student is expected to infer and construct new knowledge versus learn through the direct, explicit sharing of the new knowledge by the teacher. The NRP coded studies involving phonics to capture this continuum, comparing studies that included "systematic phonics," characterized by "a planned, sequential introduction of a set of phonics elements along with teaching and practice of these elements" (NICHD, 2000, p. 2-89), with those that included "unsystematic phonics," in which there was evidence of some phonics instruction but not in a planned, sequential fashion, and "no phonics," in which there was no evidence of any attempt to teach phonics except incidentally. In the latter programs, the instruction might focus on the teaching of whole words or might incidentally address the development of phonics skills within reading, writing, listening, and speaking activities as the need arises (NICHD, 2000). The overall effect size reported by the NRP favored systematic phonics over other forms of phonics instruction.

### Estimation of the NRP Effect Sizes: Hypothesis 1

The studies by Camilli et al. (2003, 2006) and Hammill and Swanson (2006) are empirical reassessments of the NRP results. Both essentially minimize the findings of the NRP on the efficacy of reading instruction that includes systematic phonics and therefore downplay instructional theory and pedagogy that focuses on planned, systematic instruction in the alphabetic principle for all students.

If the level of systematicity of phonics instruction is primarily responsible for the treatment effect, a dosage hypothesis may be in effect. The largest effect sizes should be associated with the strongest dose – or systematic phonics compared to the no phonics control. Smaller effect sizes should be associated with comparisons of systematic phonics with unsystematic phonics and also with comparisons of unsystematic phonics with no phonics. This simple main effects hypothesis was the initial comparison made in the NRP report and as an average, does not take into account the possibility that the instruction might interact with child characteristics; some children who are weaker in alphabetic skills may need more explicit phonics instruction (e.g., Connor, Morrison, Fishman, Schatschneider, & Underwood, 2007; Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998), whereas the degree of systematic phonics instruction may be less important for other children with better developed letter-sound knowledge. This main effect hypothesis also ignores the fact that type of instruction may

interact with school and teacher characteristics (e.g., teaching competence, time allocation; Connor et al., 2007; Foorman et al., 2006). Nonetheless, the dosage hypothesis is a useful heuristic for understanding the NRP report and its subsequent evaluations.

In calculating effect sizes, the NRP held the treatment constant, comparing the systematic phonics group from each study to any available comparison group. Some of the effect sizes represent contrasts between systematic phonics and some phonics, whereas other effect sizes represent contrasts between systematic phonics and true no phonics control groups. The  $d$  of 0.41 is the average of two groups of effects: (a) systematic phonics contrasted with some phonics, and (c) systematic phonics contrasted with no phonics. The NRP report did break down effects using type of comparison group as a moderator. The largest effects represented comparisons of systematic phonics to whole-word instruction (average  $d = 0.51$ ), and the smallest average effects represented comparisons wherein the control group had "whole language" instruction ( $d = 0.31$ ).

Camilli et al. (2003) did not accept the NRP premise that the average of different levels of phonics intervention was a meaningful comparison, focusing on study characteristics that might moderate the NRP effect size estimates. They identified two effect sizes that they felt had been miscalculated and disagreed with the inclusion of one study (Vickery, Reynolds, & Cochran, 1987) that they felt did not meet inclusion criteria and was noteworthy because it contributed eight effect sizes to the NRP database. They also identified studies in the NRP database that they felt should have been included by the NRP. Based on these modifications, Camilli et al. (2003) then recoded all of the studies from the set used by the NRP to make a different set of comparisons. They held constant the control group—always using the no phonics control and contrasting it to what they coded as "some phonics" treatments and "systematic phonics" treatments. Additionally, they coded comparisons across all groups in each study. For example, if there was a group that received a whole-word treatment, it would be compared to the no treatment (or standard practice) control group. As a result, Camilli et al. (2003) produced a much larger number of effects from these studies by coding for comparisons among all group means, which included some comparisons for which there was no phonics in either the treatment or control group. They also generated different estimates of effect size because the comparisons were adjusted to account for the presence of study moderators.

A potential limitation of these estimates is that the literature search was not designed to address the comparisons made in Camilli et al. (2003). Studies that compared no phonics and some phonics would not necessarily have been identified by the NRP search strategy. Additionally, the NRP search strategy was not designed to locate studies that compared no phonics or some phonics to a standard treatment control with systematic language activities. Finally, the NRP estimated the effect of systematic phonics instruction by taking weighted averages of the effects, both overall and within specific hypothesized moderator groups. Camilli et al. (2003) estimated the effects of systematic phonics instruction by predicting the effects via a regression equation that included recoding of the NRP studies for the amount and degree of systematicity of phonics in the treatment; whether there were "systematic language activities" in the treatment group and/or control group; and the intensity of the treatment delivery, repre-



sented as tutoring versus small-group-/classroom-level delivery. The effect size compared to the NRP overall  $d = 0.41$  was a regression weight that represented the difference between systematic phonics and some phonics controlling for the other intervention characteristics that were in the model.

Altogether, the NRP report and Camilli et al. (2003) ask different questions. The NRP question is analogous to asking about the value of receiving the intervention versus not receiving the intervention. The Camilli et al. (2003) report is analogous to asking what is the value of receiving a strong form of the intervention compared to receiving weaker forms of the intervention and relative to factors that moderate the outcomes. From our view, both questions are reasonable for intervention studies; it is worthwhile to know the average effect of supplying the intervention (i.e., the NRP report approach), but it is also worthwhile to know the incremental value of adding the intervention over and above the value of other characteristics of students, classrooms, and teachers that might affect outcomes (i.e., the Camilli et al. [2003] approach).

One would expect the estimates of effect size from Camilli et al. (2003) and the NRP to differ because the questions were different and (a) the results were based on a slightly different set of studies; (b) a different approach was used to code effects; and (c) the overall NRP effect was an unadjusted effect, whereas the Camilli et al. (2003, 2006) effects were adjusted for other study characteristics. Our first hypothesis is that when the same question is asked of the Camilli et al. (2003) data (i.e., the parameters estimated are the same), the magnitude of the effects of systematic phonics in Camilli et al. (2003, 2006) will be comparable to that of the NRP report.

### Comprehensive Approaches to Reading Instruction: Hypothesis 2

In addition to the NRP report, other consensus reports, including the National Research Council report (Snow, Burns, & Griffin, 1998) and the report of the RAND Reading Study Group (2002), have argued for comprehensive approaches to reading instruction that include literacy activities that extend beyond alphabets. This question was tested by Camilli et al. (2003) and further addressed by Camilli et al. (2006), who employed a multilevel analysis to account for the dependence of multiple effect sizes from the same studies. As discussed previously, Camilli et al. (2003) also coded for other components of instruction that were wrapped in with phonics instruction in the treatment, including systematic language activities and intensity. Note that the coding of language activities in Camilli et al. (2006) used "systematic" in a different manner from the NRP, referring primarily to the presence of literacy activities believed to facilitate fluency and comprehension and not to the degree of explicitness. These activities represented the extent to which teachers incorporated components of reading that extend beyond alphabets, emphasizing a print-rich environment, independent reading, purposeful writing, the use of invented spelling, and the use of literature to teach higher order skills. Tutoring is not really another component of reading but is a means of delivering instruction in a manner that allows for more practice at a smaller teacher-student ratio, thus increasing the intensity of instruction.

In the final models presented by Camilli et al. (2003, 2006), the individual components of the instruction all added significantly to the prediction of the effect size. Not emphasized was the implication that the overall effect size for a more comprehensive approach could be quite large. To address this question we repeated the multilevel regression analysis with modifications to the data set that permitted a more straightforward presentation of the value of adding additional activities to the reading program. Our second hypothesis is that in this stricter comparison, which allows separate empirically derived estimates of the effect of additional literacy activities on top of both some phonics and systematic phonics, effect sizes will be larger when additional literacy activities and tutoring are added to the effects of systematic phonics instruction.

### Effect Sizes in Context: Hypothesis 3

Even the critics of the NRP report concede that different studies show small effects significantly favoring systematic phonics; they disagree that these effect sizes are practically important, which is the point of departure for Hammill and Swanson (2006). Evaluating the practical utility of treatment effects is an important issue, which Hammill and Swanson addressed by translating the  $d$ s into their associated  $r$ s and  $R^2$ s and then classifying these values using an idiosyncratic variation on Cohen's suggested rules of thumb for choosing an effect size for a priori estimation of statistical power. We do not question the general heuristic labeling of effect sizes as *small* ( $r = .1$  and  $d = 0.2$ ), *medium* ( $r = .3$  and  $d = 0.5$ ), or *large* ( $r = .5$  and  $d = 0.8$ ) (Cohen, 1988) but question the use of these heuristics by Hammill and Swanson because they were never intended to be strict rules or thresholds for judging the utility of intervention effect sizes.

Cohen proposed that when conducting a power analysis in the planning stages of research, the researcher should first look to prior research in the same field to get an idea of the size of an effect that might be expected, which permits an estimate of the sample size that will be required to minimize the risks of a Type II error (in this context, identifying a treatment as ineffective when in fact it was effective). If the researcher is not able to obtain a numeric value to use in the power analysis calculations, Cohen (1988) proposed the use of the now familiar effect size heuristics but warned against their misuse and abuse. He stated:

The author proposes *as a convention*, [effect size] values to serve as operational definitions of the qualitative adjectives "small", "medium", and "large." This is an operation fraught with many dangers: The definitions are arbitrary, such qualitative concepts as "large" are sometimes understood as absolute, sometimes as relative; and thus, they run the risk of being misunderstood. (p. 12)

Cohen (1988) also insisted that the importance of an effect could not be discerned from its size independently of its context, indicating that "the *meaning* of any given [effect size] is, in the final analysis, a function of the context in which it is embedded" (p. 535). Small effect sizes for important outcomes may have significant implications in a practical context (Trusty, Thompson, & Petrocelli, 2004). Furthermore, small effects in ongoing processes, such as the development of key skills underlying reading, may accrue over time to become moderate to large effects (Prentice & Miller, 1992). Conversely, large effect sizes may be trivial within practical contexts because the spuriously large effect was the result



of method variance and response biases (Thorndike, 1997) and from specification error and outliers (Pedhazur, 1982). As Cohen (p. 535) stated,

"only 50% of the variance" may be as valid a formulation in one context as "only 1% of the variance" is in another, and conversely, "as much as 1% of the variance" is, in principle, no less valid a formulation than "as much as 50% of the variance."

Thus, one-size-fits-all rules of thumb may not help to interpret an effect size in the absence of the context in which it is to be operationalized (Vacha-Haase & Thompson, 2004).

Considering the caveats provided by Cohen (1988), the labeling of effect sizes can serve as an aid to communication, in the sense that a large effect from a high-quality study is nearly always "better" than a small or medium effect. Although difficulty in making decisions about the practical impact of a given effect size is less likely in the case of an effect size with a large value (e.g., in the range of about  $d = 0.80$ ), even here one might envision a hypothetical situation in which, given a significant cost to obtain the result, an effect size of this absolute magnitude might carry little practical meaning. For example, individual tutoring may have a larger effect size than classroom-based instruction, but providing 1:1 instruction to every single student in public schools carries an unrealistic cost, not to mention the logistical difficulties of training sufficient numbers of teachers, rebuilding schools to support tutoring, and so on. However, as effect sizes become progressively lower in absolute value, an evaluation of contextual factors is more likely to be required, which is more complicated than simply comparing an effect size to some benchmark. A consideration of costs and benefits may be required. Although labels can be used with effect sizes, the utility of the effect under consideration is not synonymous with its label. For example, although the translation of *small* into  $d = 0.2$  could be useful when planning research, the translation of *small* into *of little utility and of little practical importance* is a much more tenuous proposition and is likely to be biased without a consideration of context.

A clear example of such an effect is the well-known 5-year randomized study that led to the recommendation to utilize aspirin in the prevention of heart attack (Steering Committee of the Physicians' Health Study Research Group, 1989). The absolute value of the effect size from this study was  $d = 0.07$ , which is *negligible* under heuristic guidelines of standard deviation units and certainly unimpressive in terms of correlation ( $r = .034$ ) or in terms of proportion of variance ( $R^2 = .001$ ). However, given the low base rate of heart attack (1.33% in that study), the effect size of  $d = 0.07$  means that nearly twice as many individuals in the placebo group suffered a heart attack relative to individuals in the aspirin condition. In the context of the low base rate and high costs of heart attacks and the generally low cost of aspirin, this effect size moves from unimpressive to powerful.

The average effect sizes of different educational interventions are larger than that in the aspirin example. Lipsey and Wilson (1993) reported that the average educational intervention had an effect size of  $d = 0.34$ , or  $r = .168$ , which is  $R^2$  of .028. In another large meta-analysis, Swanson, Hoskyn, and Lee (1999) reported an average weighted effect size of  $d = 0.56$  for the effects of educational interventions in people identified with learning disabilities across a wide variety of outcome measures. The latter effect size is larger than the average effect for phonics in the NRP report,

which in turn is larger than that of Lipsey and Wilson. Potential explanations for the variability of these results include differences in the interventions; the populations, which differed by age and subgroup; as well as methodological quality; research design; and so on. The Lipsey and Wilson meta-analysis involved a variety of educational interventions in kindergarten through Grade 12 and college. In contrast, the effect size reported by Swanson et al. was for individuals with learning disabilities in kindergarten through Grade 12. The NRP effect size of  $d = 0.41$  was for phonics interventions in kindergarten through Grade 6 over all subgroups. However, note that for poor readers, the NRP reported an effect size of  $d = 0.98$  for decoding regular words and  $d = 0.67$  for decoding pseudowords in kindergarten through Grade 2 and  $d = 0.49$  for decoding regular words and  $d = 0.52$  for decoding pseudowords in Grades 2 through 6. These latter values are especially comparable to the  $d = 0.57$  reported by Swanson et al. for the effect of intervention in groups with learning disabilities on word recognition. Obviously none of these meta-analyses were exact replications of one another. The amount of variation in these estimates is consistent with the possible effects of the moderators mentioned previously as well as expected sampling error. The next step, however, is not to compare the obtained effect to a one-size-fits-all rule of thumb or convention, but to use these estimated effects to determine the utility of implementing various types of interventions by taking into account contextual factors.

In the case of interventions designed to improve reading performance, identifying factors such as the base rate of struggling readers and the cost of interventions as well as the cost of not providing an intervention (e.g., the cost that might be represented by dropping out of high school) can assist in contextualizing the effect sizes found by either the NRP study or the Camilli et al. (2003, 2006) analyses. To address this question, we used the effect size estimates from the NRP report and different estimates of the incidence of reading difficulties and a variant of a binomial effect size display (Rosenthal & Rubin, 1982) to evaluate the importance of context in interpreting effect size data. Consistent with Cohen (1988), our third hypothesis is that interventions with effect sizes as small as those identified by Hammill and Swanson (2006) for phonics instruction could significantly reduce the number of children with reading problems depending on the base rate used to estimate the incidence of reading difficulties and the effect size associated with different interventions.

## Hypothesis 1

### Method

**Database.** To address the first hypothesis concerning the impact of study parameters on the discrepancies in the NRP report and in Camilli et al. (2003), we relied on the corpus of studies identified by the NRP and recoded by Camilli et al. (2003, 2006). These studies and the database of effect sizes and codes were accessed from the online journal *Education Policy Archives Analysis* Web site in which Camilli et al. (2003) originally appeared (<http://epaa.asu.edu/epaa/v11n15>). Although several questions were asked in the section of the NRP report addressing phonics, the first question was the target of both Camilli et al. (2006) and Hammill and Swanson (2006): "Does systematic phonics instruction help children learn to read more effectively than nonsystem-

atic phonics instruction or instruction teaching no phonics?" (NICHD, 2000, p. 2-92). The primary criterion for including studies into this meta-analysis was stated in the NRP methodology: "Studies had to provide data testing the hypothesis that systematic phonics instruction improves reading performance more than instruction providing unsystematic phonics or no phonics instruction." (NICHD, 2000, p. 2-90). This criterion requires that a study include a comparison of systematic phonics instruction versus any other control condition, which is what the  $d = 0.41$  in the NRP report represents. Some of these control conditions will have no phonics and others will have unsystematic phonics. With this strategy and other criteria, the NRP screened 75 studies representing randomized or quasi-experimental designs with treatment and control groups, with 38 represented in the final database. Camilli et al. (2003) included substantially the same set of studies included in the NRP meta-analysis, removing one study and adding three from the corpus of studies identified by the NRP (no new search was conducted).

**Predictors.** The dependent variable in Camilli et al. (2003) was the effect size, and the independent variables were coded vectors that represented many other characteristics of the intervention given in each study. After a stepwise regression procedure, their final model included as predictors the amount and degree of systematicity of phonics in the treatment; the presence of systematic language activities in the treatment group and/or control group; and the intensity of the treatment delivery, represented as tutoring versus small-group-/ classroom-level delivery.

**Procedures.** We reformulated the results from the two studies within the same framework. To further understanding of the dosage hypothesis and what it means for comparing effect size estimates from the NRP report and Camilli et al. (2003), consider Figure 1. The horizontal axis represents performance on some reading outcome measure. We hypothesized that the average performance associated with no phonics would be at the far left (lowest performance), the average performance for systematic phonics would be on the far right (highest performance), and the average performance associated with unsystematic or some phonics would be in between these two extremes. The line segments connecting the three groups represent the distance in effect size units between each pair. Line segment *a* represents the average effect between systematic phonics and a no phonics control group,

line segment *b* represents the average effect between some phonics and a no phonics control group, and line segment *c* represents the difference between systematic phonics and some phonics groups. We can use this generic framework to highlight the different elements that are being considered in the NRP report and in the Camilli et al. (2003, 2006) reports and to demonstrate why we would not expect the numbers they report to be the same, even in the context of the same studies.

The NRP asked about the effects of systematic phonics instruction. Therefore, they coded some studies that compared systematic phonics to a no phonics control (represented by line segment *a*) and other studies that compared systematic phonics to a some phonics control (line segment *c*). They then averaged the set of *a* and *c* effects to get  $d = 0.41$ . Note that the size of this average depends on the average size of *a*, the average size of *c*, and the number of studies providing estimates of *a* and *c* found in the literature. If, consistent with our model, *as* are systematically larger than *cs* and there are more of them, the overall average will be larger than if the proportion of *as* and *cs* had been reversed. The NRP also included a test of the homogeneity of the distribution of effects and found that they were not homogenous. When effects are not homogenous, presenting the overall effect is only Step 1, followed by presentation of the effects within moderator groups. As a result, the NRP presented the average effects within many potential moderator groups, including one breakdown that showed that the effect size differed depending on the type of control group.

Camilli et al. (2003) recoded all of the effects to arrive at a different set of comparisons. They made all comparisons against a no phonics control, thus estimating the quantities *a* and *b* in Figure 1. Note that they also coded estimates of the difference between groups where neither of the groups received any phonics instruction. An example would be a comparison between a treatment group receiving whole-word instruction and a control group receiving standard instruction. This contrast cannot be represented in Figure 1, which represents the continuum of the phonics treatment effects. In studies that included a systematic phonics group, a some phonics group, and a no phonics control group, Camilli et al. (2003) coded both difference *a* and difference *b*, thus obtaining more comparisons from the same set of studies than the NRP analysis. They then analyzed the coded effects in a regression

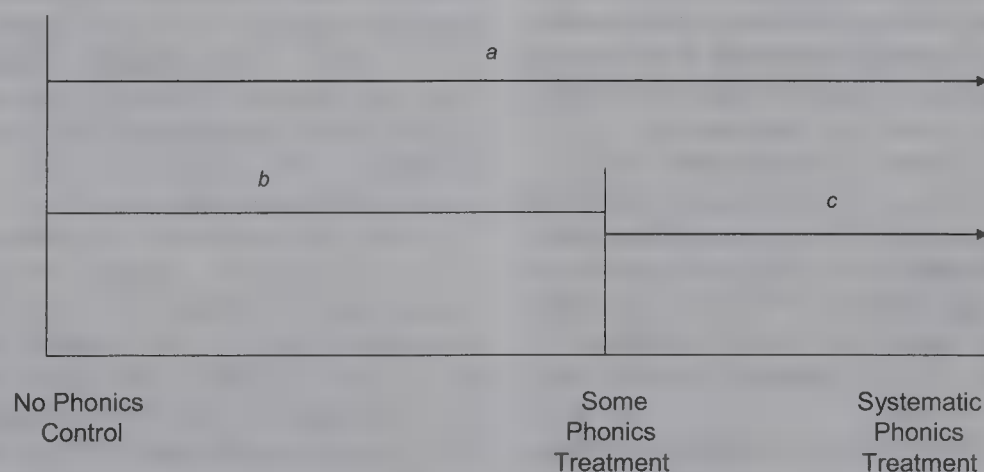


Figure 1. Model for comparisons of effect sizes from the National Reading Panel report and Camilli et al. (2003, 2006).



analysis to estimate and test the  $c$  difference (i.e., the difference between systematic phonics and some phonics).

The conceptual difference between the two approaches is that the NRP estimated the average of a mixed set of effects ( $a$  and  $c$ ) and then tested the distribution of these effects to determine if they were homogenous. Camilli et al. (2003) estimated  $c$ , the average difference between the estimates of effects  $a$  and  $b$ , and tested whether there was a significant difference when the set of effects was divided along the some phonics–systematic phonics dimension. The overall average, or the NRP estimate, is almost certainly going to be larger than  $c$  regardless of the mix of  $a$ s and  $c$ s, provided there is at least one  $a$  effect in the group of estimates. We used the model in Figure 1 as a framework for comparing estimates of effect size in Camilli et al. (2003) and the NRP report.

## Results

Camilli et al. (2003) provided several different estimates for  $c$ . The first was presented in the context of their Table 3, in which the average univariate effect for comparisons of systematic phonics with no phonics controls was  $d = 0.514$  and the average comparison of some phonics with no phonics controls was  $d = 0.243$ . To put this into the context of Figure 1, the former was their estimate of the average  $a$  effect; the latter was their estimate of the  $b$  effect. We can calculate the difference between the two to arrive at an estimate of  $c = .27$ . Camilli et al. (2003) pointed out that their estimate of the effect of systematic phonics was about 30% smaller than the effect reported by the NRP. They failed to underscore the fact that .27 is only an estimate of the average  $c$  effect and not, like the NRP average, an estimate of the average of the mixed  $a$  and  $c$  effects, which Figure 1 shows must be larger. We do not know how many  $a$  and  $c$  effects were coded by the NRP because we based our reanalysis on the Camilli et al. (2003) data set, which did not contain this information. However, if we make the simplifying assumption that there are equal numbers of  $a$  and  $c$  effects and use these estimates of  $a$  and  $c$  (.514 and .27), our estimate of the average effect is  $(.514 + .27) / 2$  or  $d = 0.39$ , which is remarkably close to the value of  $d = 0.41$  obtained in the NRP analysis. Thus, when we estimate the same parameter, the results of the two coding rubrics converge despite slight differences in the corpus of studies.

Other estimates of the  $c$  effect can be obtained from Camilli et al. (2003). Specifically, their Tables 6 and 7 display the parameter estimates from their final regression equations after stepwise analysis, in which the effect sizes are predicted from a number of study characteristics, including the amount of systematic language instruction in the treatment, the amount of systematic language instruction in the control group, and whether tutors were used to deliver instruction versus small groups or whole classrooms. These two sets of results differed in the weights that were applied in the analysis. The results from their Table 6 used weights ( $WGT$ ) that gave equal representation to each study, where  $k$  represents the number of records contributed to the database by each study:  $WGT1 = 1 / k$ .

The results from their Table 7 used compromise weights ( $WGT3$ ), or weights that resulted from multiplying equal representation weights with optimal weights (Hedges & Olkin, 1985, p. 110), which are a function of the sample size and effect size from

each study. See Equations 1 and 2 and Camilli et al.'s (2003) Table 4 for a description of these weights:

$$WGT3 = WGT1 \times WGT2 \quad (1)$$

where

$$WGT2 = \left( \frac{n_{TOT}}{n_T n_C} + \frac{d^2}{2(n_{TOT} - 2)} \right)^{-1} \quad (2)$$

For both analyses, the regression weight associated with  $TP2$  is the parameter estimate of interest. This parameter represents the difference between (a) the effects that compare systematic phonics with a no phonics control and (b) those that compare some phonics with a no phonics control, which, in the context of Figure 1, is an estimate of  $c$ . However, in these models, the effect size estimate has been adjusted for the presence of other study characteristics; the systematic phonics effect has been corrected for the presence of systematic language activities in the same treatment or for treatment delivery through tutoring. The effect is to make the average of the  $a$  effects and the  $b$  effects closer to each other. When equal representation weighting is used (i.e.,  $WGT1$ ), the best estimate of  $c$  is .241; when compromise weighting is used (i.e.,  $WGT3$ ), the best estimate of  $c$  is .188. These estimates are lower than the NRP estimate but should not be directly compared to it because the NRP estimate is an average of the  $a$  and  $c$  effects.

In the same vein, Camilli et al. (2006) presented a third estimate of  $c$  in the context of a multilevel model that permitted control for the dependencies among effects from the same study. The parameter estimates from this model indicated that the effect of moving from some phonics to systematic phonics was  $d = 0.123$  when other study characteristics and the dependency of effects within studies were controlled. Again, this is an estimate of the  $c$  effect adjusted for study moderators and cannot (and should not) be directly compared to the NRP unadjusted average of a mixed set of  $a$  and  $c$  effects of  $d = 0.41$ .

Another approach to estimating the effect size obtained by the NRP from the Camilli et al. (2003) recoded and reanalyzed data would be to use the parameter estimates from the regression model in which compromise weighting is used. Our best estimate of the  $a$  effects may be obtained by adding the constant from this model (.349) to half of the regression weight (because of the effect coding that was used) for  $TP2$ , or .188. We obtain an estimate for the effects of systematic phonics when compared to a no phonics control of  $.349 + .094$ , or  $d = 0.443$ . We can then use the parameter estimate for  $TP2$ , or .188, as our best estimate of the size of the  $c$  effect, or the difference between groups that received systematic phonics and those that received some phonics. If we assume that we had equal numbers of  $a$  effects and  $c$  effects to average, the best estimate of the overall average effect (without weighting by sample size) would be  $d = 0.316$ , or the average of .443 and .188. If there were more studies that evaluated the effect of systematic phonics compared to a no phonics control, the average would be closer to .443; if more studies evaluated the effect of systematic phonics versus a some phonics control, the average would be closer to  $d = 0.188$ . Although this average  $d = 0.316$  is smaller than the  $d = 0.41$  reported by the NRP, remember that this estimate represents an effect that has been adjusted for other study variables, such as the inclusion of language activities and tutors, so it should be smaller because the NRP estimates did not adjust for these moderators.



## Hypothesis 2

## Method

**Database.** To test the second hypothesis that combining different instructional activities was, on average, more effective, we used the Camilli et al. (2003) selection of studies and their coding of the NRP database and ran the regression model using the multilevel approach in Camilli et al. (2006). The analysis took into account the clustering of the data or the nonindependence of effects from the same study. It was not our intent to carry out a full multilevel model analysis with estimates of both fixed and random effects, but to replicate the regression analyses carried out by Camilli et al. (2006) on the modified database. Camilli et al. (2006) extended the work done in the 2003 reanalysis by recoding all of the studies for the level of systematic language activities in both the treatment and control groups using a 3-point rubric where a code of 0 indicated *no literacy activities*, 1 indicated *some literacy activities*, and 2 indicated *multiple language activities*.

**Procedures.** We deleted 25 effects from the total of 224 effects used in their analysis where there was neither some phonics nor systematic phonics instruction in the treatment group because we were interested in modeling treatment effects for phonics instruction and not for treatments that included no phonics. To facilitate interpretation of the regression parameter estimates, we recoded the *TP* variable into *TP\_di*, where 0 represented *some phonics instruction* and 1 represented *systematic phonics instruction*. We agreed that controlling for the presence of other instructional characteristics in both the treatment and control groups might help explain some of the heterogeneity found in the NRP study and would also potentially yield a less biased average effect.

Other than these two modifications, we set up the multilevel model analysis in the same way as Camilli et al. (2006) using the SPSS code for their analysis, which was available from the Web site for Camilli et al. (2003). We translated their code into SAS Proc Mixed code and then reran their analyses on their original variables and the full set of effects to verify that we were running the same multilevel model. We then ran this model on the smaller, recoded data set (from which the 25 superfluous effects had been deleted) and calculated the predicted effect size values from the resulting model for all combinations of our predictor variables: the amount and systematicity of phonics instruction in the treatment group (*TP\_di*), the amount and systematicity of language activities also present in the treatment group (*TL2*), the amount and systematicity of language activities in the control group (*CL2*), and whether or not the intervention was delivered one on one versus in small groups or classrooms (*Tutor*). The coding for *TL2*, *CL2*, and *Tutor* were included in the published database and were not changed.

## Results

The estimates of the fixed effects from our multilevel regression analysis are presented in Table 1. All of the treatment characteristics included in the model significantly predicted unique variance in the effect sizes. To evaluate the practical impact of these various treatments options alone and in combination, we combined, as in any standard regression-based prediction model, the coded values for the predictors with the regression weights (see Table 1) to calculate predicted mean effects for each combination of predictor

Table 1

*Fixed Effects From the Multilevel Model Modified From Camilli et al. (2006)*

Effect	Estimate	SE	<i>t</i>	<i>p</i>
Intercept	.306	.106	2.88	.007
<i>TP_di</i> <sup>a</sup>	.183	.089	2.06	.042
<i>TL2</i> <sup>b</sup>	.211	.106	1.98	.049
<i>CL2</i> <sup>c</sup>	-.213	.079	-2.71	.008
<i>Tutor</i> <sup>d</sup>	.424	.158	2.69	.008

Note. *df* = 36 for intercept; *df* = 158 for each of the other effects.

<sup>a</sup> Phonics in the treatment group: 0 = *unsystematic*, 1 = *systematic*.

<sup>b</sup> Language activities in the treatment group: 0 = *none*, 1 = *some*, 2 = *systematic*. <sup>c</sup> Language activities in the control group: 0 = *none*, 1 = *some*, 2 = *systematic*. <sup>d</sup> Intensity of instruction: 0 = *classroom- and small-group level*, 1 = *one on one*.

values. We then sorted the predicted values of the effect sizes from smallest to largest and present the mean predicted effect sizes along with the study characteristic codes in Table 2. Note that these analyses are based on the literature search by the NRP and the recoding of data by Camilli et al. (2006) and are therefore restricted to this body of studies.

Each row of Table 2 represents a potential combination of treatments and the expected effect size for that combination of treatments based on existing data. For example, the lowest *d* (-0.121), which may be found in the first row of the table, is the predicted value of *d* when there is some phonics in the treatment group, no language activities in the treatment group, no tutor, and systematic language activities in the control group. Under this condition a negative effect size favoring the control group is not surprising given the unique positive effect of systematic language activities that, in this comparison, are present in the control group but not the treatment group. Next, note the line in Table 2 in which all of the predictor values are 0 and the predicted effect size is *d* = 0.306. This line represents some phonics in the treatment group, no language activities in the control or treatment groups, and no tutoring. The value of predicted *d* for this scenario is equal to the intercept from our model. If we then move to the line where the *TP\_di* variable is 1 and all other predictors are 0, the predicted value of *d* is 0.488. Thus, if we compare a treatment group that receives systematic phonics, no tutors, and no additional language activities to a control group that also has no phonics or systematic language activities and is taught in small groups or whole classrooms, we can hypothesize that the treatment group would perform .488 standard deviations higher than the control group on the outcome assessment. This effect is simply the intercept plus the effect of moving one unit on the predictor while holding the other predictors constant.

Next, examine the row in Table 2 in which there is systematic phonics instruction plus 1:1 tutoring. The predicted effect is *d* = 0.913, a very large effect. If a practitioner already had a systematic phonics program in place, this table could be used to get an idea of the potential effect of adding tutoring or additional literacy components. In the first case, the hypothesized improvement would be approximately .425, or the difference between .913 and .488, not coincidentally the value of the beta weight for tutoring in this analysis. The hypothesized effect of adding both tutoring and systematic language instruction to an existing systematic phonics



Table 2  
*Predicted Values Based on the Parameter Estimates From Table 1 and Coded Values for the Predictors*

Predicted value	TP_di <sup>a</sup>	TL2 <sup>b</sup>	CL2 <sup>c</sup>	Tutor <sup>d</sup>	No. of effects	k (No. of studies)
<b>-0.121</b>	<b>0</b>	<b>0</b>	<b>2</b>	<b>0</b>	<b>12</b>	<b>3</b>
<b>0.062</b>	<b>1</b>	<b>0</b>	<b>2</b>	<b>0</b>	<b>12</b>	<b>4</b>
<b>0.090</b>	<b>0</b>	<b>1</b>	<b>2</b>	<b>0</b>	<b>7</b>	<b>1</b>
0.092	0	0	1	0		
0.273	1	1	2	0		
0.275	1	0	1	0		
0.301	0	2	2	0		
0.303	0	1	1	0		
0.305	0	0	2	1		
<b>0.306</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>30</b>	<b>6</b>
0.484	1	2	2	0		
0.486	1	1	1	0		
0.487	1	0	2	1		
<b>0.488</b>	<b>1</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>85</b>	<b>17</b>
0.514	0	2	1	0		
0.516	0	1	2	1		
0.517	0	1	0	0		
0.518	0	0	1	1		
0.697	1	2	1	0		
0.698	1	1	2	1		
<b>0.700</b>	<b>1</b>	<b>0</b>	<b>1</b>	<b>1</b>	<b>11</b>	<b>2</b>
0.727	0	2	2	1		
0.728	0	2	0	0		
0.729	0	1	1	1		
<b>0.731</b>	<b>0</b>	<b>0</b>	<b>0</b>	<b>1</b>	<b>7</b>	<b>2</b>
<b>0.909</b>	<b>1</b>	<b>2</b>	<b>2</b>	<b>1</b>	<b>3</b>	<b>1</b>
<b>0.910</b>	<b>1</b>	<b>2</b>	<b>0</b>	<b>0</b>	<b>3</b>	<b>1</b>
<b>0.911</b>	<b>1</b>	<b>1</b>	<b>1</b>	<b>1</b>	<b>4</b>	<b>1</b>
<b>0.913</b>	<b>1</b>	<b>0</b>	<b>0</b>	<b>1</b>	<b>17</b>	<b>5</b>
0.940	0	2	1	1		
0.942	0	1	0	1		
1.122	1	2	1	1		
1.124	1	1	0	1		
<b>1.153</b>	<b>0</b>	<b>2</b>	<b>0</b>	<b>1</b>	<b>4</b>	<b>1</b>
<b>1.335</b>	<b>1</b>	<b>2</b>	<b>0</b>	<b>1</b>	<b>4</b>	<b>1</b>

Note. Boldface indicates actual data from the meta-analysis. Other rows are extrapolations based on the regression equation.

<sup>a</sup> Phonics in the treatment group: 0 = *some*, 1 = *systematic*. <sup>b</sup> Language activities in the treatment group: 0 = *none*, 1 = *some*, 2 = *systematic*. <sup>c</sup> Language activities in the control group: 0 = *none*, 1 = *some*, 2 = *systematic*.

<sup>d</sup> Intensity of instruction: 0 = *classroom- and small-group level*, 1 = *one on one*.

program based on the body of studies in the NRP report as recoded by Camilli et al. (2003, 2006) is  $d = 1.34$ .

Note that the nonbolded rows in Table 2 represent combinations of the predictor variables for which there are no studies in the extant literature from which to estimate effects. The predicted  $d$  in this case is an interpolation based on the data in hand. We have bolded the rows that represent predicted effects based on actual effects in the meta-analysis and have included the number of effects and the number of studies for each of these rows. Obviously, we would have more confidence in estimating effects based on more substantial data than on only a few studies or on interpolation.

The estimates from Table 2 can be used to generate hypotheses of anticipated results that need testing in an experimental study. Caution should be exercised when comparing the effects of systematic phonics instruction to those for tutoring or language activities from Table 2. The set of studies identified by the NRP was the result of a search methodology designed to find all possible

studies that evaluated systematic phonics instruction using a randomized design or a high-quality quasi-experiment. This search was not designed to locate and include all possible studies evaluating the effect of using tutors to deliver instruction or of using language activities. Confidence in the magnitude of the systematic phonics effect is stronger because the NRP sought and identified the population of published studies.<sup>1</sup> The sample of tutoring and language activity effects contained within these studies is probably not the population of all such published effects and is also not a random sample from the population of all such studies; as a result, it is subject to selection biases. The effect on our results is that we

<sup>1</sup> Meta-analysis of the effect of group size do not support the greater efficacy of small-group tutoring in sizes of 1:1 versus 3:1 (Elbaum, Vaughn, Hughes, & Moody, 2000). Camilli et al. (2003, 2006) and the NRP compared 1:1 tutoring with the combined effects of small-group and classroom instruction.

should have less confidence in the precise magnitude of these effects or our ability to estimate the additive models laid out in Table 2. These cautions also apply to the conclusions reached by Camilli et al. (2003, 2006) and Hammill and Swanson (2006) to the extent that they are based on study characteristics other than the effect of systematic phonics.

### Hypothesis 3

#### Method

**Procedures.** Using the effect size estimates from the NRP report and different estimates of the incidence of reading difficulties, we used a variant of a binomial effect size display (Rosenthal & Rubin, 1982) to evaluate the practical significance of effect sizes of different magnitudes under different base rates of struggling readers. In this variant we used realistic base rates rather than the 50% base rate in the original formulation of binomial effect size displays.

**Statistical procedures.** To create these displays, we took advantage of the fact that the relation between two dichotomized variables may be summarized with the phi coefficient, which is equivalent to the correlation coefficient effect sizes used by Hammill and Swanson (2006). We created a  $2 \times 2$  table to represent the relation between a phenomenon of interest with two levels (high school dropout vs. high school completion) and a treatment with two levels (treatment vs. control). We placed the existing base rate for the phenomenon into the control level of the treatment and then, for a given level of phi, calculated the frequency of the phenomenon in the treatment group required to obtain that level of phi. Table 3 contains a table of symbolic frequencies to be used in conjunction with the formula for phi to solve for the unknown frequency for a given base rate and given phi:

$$\Phi = (ad - bc) / \sqrt{efgh}.$$

Using  $a$  through  $h$  in this equation, if we assume a value for  $c$  based on our knowledge of the base rate of a given phenomenon, and set  $g$  and  $h$  equal to each other, we can solve for the value of  $d$  that would correspond to any given phi value. For our examples, we chose total frequencies within the treatment and control to be equal ( $n = 10,000$  each) and large enough to allow us to calculate frequencies in the treatment group in whole numbers. We used our assumed base rate to obtain the frequencies  $a$  and  $c$ . If the base rate of high school noncompletion were 10%, we would set  $a$  equal to 9,000 and  $c$  equal to 1,000 and solve for the value of  $d$  associated with a given level of phi.

**Hypothetical scenarios.** We created three hypothetical examples addressing the potential impact of even small effects within

the context of reading instruction. In Scenario 1, we used the rate of at-risk readers in early intervention programs, commonly (but somewhat arbitrarily) estimated at 20% (Torgesen, 2000), as the characteristic we would like to change. In Scenario 2, we reduced the effect size in Scenario 1 and lowered the base rate to the high school noncompletion rate of 10% (Laird, DeBell, & Chapman, 2006) as the phenomenon we would like to reduce. In Scenario 3, we maintained the small effect size in Scenario 2 but increased the base rate to represent the number of struggling readers in the United States, which was cited as up to 40% by Hammill and Swanson (2006; see Shaywitz, 2004).

#### Results

In the first scenario, consider  $d = 0.48$ ,  $r = .23$ , an effect size called "small" by Hammill and Swanson (2006). This effect was chosen because it was the effect for typical achievers in first grade from the NRP report. In Scenario 1, assuming a sample of 20,000 (half treatment, half control) and a base rate of 20% of at-risk readers who participate in a reading program with a predicted  $d = 0.48$ ,  $r = .23$ , Table 4 shows a cross-tabulation of treatment and at-risk classifications. An intervention that results in a phi coefficient of .23 between the treatment and classification is consistent with an expected frequency of at-risk readers in the intervention group of 5%, a substantial decrease from 20%.

Similarly, consider an intervention that yields an effect size  $d = 0.23$ ,  $r = .11$ , with respect to the high school dropout rate (see Table 4). We selected this effect size, the smallest significant effect reported in the NRP meta-analysis, because not only was it labeled small by Hammill and Swanson (2006), but it is also close to the Cohen (1988) heuristic for a small effect size. Assuming a 10% high school dropout rate without an intervention (Laird et al., 2006), an intervention that correlates with dropout status at the  $r = .11$ ,  $R^2 = .012$ , level could result in a reduction in the dropout rate from 10% to 4.5%.

Finally, consider an intervention for which we expect a small effect of  $d = 0.23$ ,  $r = .11$ , and set our base rate for struggling readers at 40%. Adjusting the base rate so that it is higher should reduce the potency of an intervention with a small effect size. Even with an intervention with this small effect size, we could hypothesize a reduction in incidence of poor readers from 40% to 30% (see Table 4).

### General Discussion

There are three major conclusions from our assessment of Camilli et al. (2003, 2006) and Hammill and Swanson (2006). First, in terms of the first hypothesis, the conclusions of the NRP report are not contradicted by the two reanalyses of Camilli et al. (2003, 2006). The comparisons by Camilli et al. (2003, 2006) ask questions that are different from the primary question asked by the NRP, but the results of the two sets of analyses can be reconstructed to yield comparable effect sizes for the effects of systematic phonics versus either unsystematic phonics or no phonics controls when the same study parameters are estimated. When the effect sizes have been adjusted for study moderators, as in Camilli et al. (2003, 2006), the estimates are expected to be lower than the NRP estimate because the other treatment characteristics tend to improve performance (i.e., are positively related to the effect size)

Table 3  
Example of Symbolic Frequencies to Be Used With the Formula for Phi to Solve for the Unknown Frequency for a Given Base Rate and Phi

Variable	Control	Treatment	Total
High school completion	$a$	$b$	$e$
Noncompletion	$c$	$d$	$f$
Total	$g$	$h$	$a + b + c + d$



Table 4  
*Binomial Effect Size Display for 10,000 Students in Scenarios 1, 2, and 3 After Intervention*

Group	Classification	No intervention	With intervention
Scenario 1			
At-risk readers <sup>a</sup>	No	8,000	9,500
	Yes	2,000	500
Scenario 2			
High school dropouts <sup>b</sup>	No	9,000	9,550
	Yes	1,000	450
Scenario 3			
Struggling readers <sup>c</sup>	No	6,000	7,000
	Yes	4,000	3,000

*Note.* Scenario 1: base rate = .20,  $d = 0.48$ ,  $r = .23$ ; Scenario 2: base rate = .10,  $d = 0.23$ ,  $r = .11$ ; Scenario 3: base rate = .40,  $d = 0.23$ ,  $r = .11$ .  
<sup>a</sup> Incidence of at-risk readers reduced from 20% to 5%. <sup>b</sup> Incidence of high school dropouts reduced from 10% to 4.5%. <sup>c</sup> Incidence of struggling readers reduced from 40% to 30%.

and because the vectors of study characteristics are not orthogonal to one another. If the estimates are directly compared, the key is to estimate the same effect size across the different studies. An effect size is always a comparison of different conditions with one another (see Figure 1). The NRP and Camilli et al. (2003, 2006) estimated effects from different comparisons.

Second, Camilli et al. (2003, 2006) showed that both phonics and language instruction, as well as tutoring, impacted reading outcomes. They questioned whether it was accurate to assume that phonics instruction was the core of many of the programs represented in the NRP database, especially interventions in the control groups. These findings support the NRP contention that reading programs need to be comprehensive by providing estimates of the unique contributions of different factors that moderate the impact of phonics. To get larger effect sizes, reading instruction must be comprehensive and contain literacy components beyond phonics, especially when the diversity of students in classrooms and local instructional contexts are considered. Thus, in terms of the second hypothesis, reanalyses of the NRP report show that larger effects may occur in association with combining different components of reading instruction. Within the modified NRP corpus of studies, systematic phonics was associated with larger effects than no or some phonics, but the effects tended to be small to moderate and variable across studies, and much of the explainable variability in effects (i.e., that not associated with sampling error) was not accounted for by the intervention characteristics included in the model. The addition of literacy components that presumably support fluency and comprehension, as well as tutoring, was associated with larger effects. However, larger effects were associated with systematic phonics, regardless of the levels of systematic language activities and tutoring. Within the modified NRP corpus of studies, the largest effects were associated with the combination of systematic phonics with additional language and literacy activities and one-on-one tutoring.

Third, whether we characterize  $d = 0.41$  as moderate or small, the evaluation of Hypothesis 3 shows that even small effect sizes may be of sufficient magnitude that they could be associated with significant reductions in the incidence of reading problems. Hammill and Swanson (2006) utilized Cohen's rules of thumb for planning studies as thresholds for determining the practical significance of effect sizes but did not take into account context and base rates in their dismissal of small effect sizes, essentially committing a Type II error. Although they were correct in suggesting that additional variability is unexplained, it is an inappropriate extrapolation to suggest that the amount of explained variability is not practically significant. In Scenario 1, with a medium effect and a medium small base rate, the hypothetical intervention reduced the incidence of struggling readers from .20 to .05, for a reduction of 75%. In Scenario 2, with a small effect and a lower base rate than Scenario 1, the hypothetical intervention reduced the incidence from .10 to .045, for a reduction in struggling readers of 55%. In Scenario 3, with the same small effect as Scenario 2 and a high base rate, the given effect size was associated with a reduction in struggling readers from .40 to .30, or 25%. These examples show that the impact of even small effect sizes may be practically important, especially when coupled with low base rates of the phenomenon of interest. The next step in placing these effects into a context involves computing the costs associated with delivering a given intervention with the benefits expected from moving some number of individuals from one category to another and comparing these costs with expected benefits. This step is outside of the bounds of this article, and, in fact, because costs and benefits are context dependent, the practical significance of a given effect size might be decided on a location-by-location basis. However, because reading instruction is routinely provided to students in schools, the costs in changing instructional emphases should be relatively small compared to the overall costs already in place for teaching children to read. The effect size and the base rates used in our examples are comparable to situations that exist with different school settings and interventions.

Altogether, these results support approaches to reading instruction that are more comprehensive and, for alphabetics, approaches that are more explicit and in which the knowledge is directly shared relative to those in which knowledge must be inferred by students. Whether these principles extend beyond just the effects of phonics instruction cannot be established from the NRP report or Camilli et al. (2003, 2006), although other reviews support more explicitness for fluency and comprehension (Pressley, 2005). In reaching this conclusion, we note that these pedagogical principles exist on a continuum and should not be dichotomized. In examining this continuum for instruction involving the alphabetic principle, it may be that the more important component is explicitness and the deliberate attempt to instruct the child as opposed to a scripted approach to phonics, especially if the child is at risk for reading difficulties or is struggling to learn to read. Indeed, both the NRP and Camilli et al. (2006) concur in estimating larger effects of systematic phonics for students who are struggling readers, findings supported by recent experimental studies that formally manipulate explicit instruction in relation to child characteristics. For example, Connor et al. (2007) found that more time in phonics instruction is beneficial to students weak in alphabetic knowledge; conversely, more time on comprehension instruction



leads to better outcomes in students weak in vocabulary instruction.

To illustrate the difference in scripted versus explicit, Torgesen et al. (2001) compared the efficacy of a highly scripted reading program with a clearly defined scope and sequence with an approach that taught the alphabetic principle explicitly but spent more time reading and writing in context. There were no significant differences in outcomes for a group of elementary school children with severe reading disabilities (see also Wise, Ring, & Olson, 2000). Mathes et al. (2005) compared two comprehensive small-group tutorial interventions based on (a) a direct instruction model with a scripted lesson plan and well-developed scope and sequence with use of decodable text; and (b) a guided reading intervention in which instruction in the alphabetic principle was explicit (i.e., based on a plan for introducing phonics elements and in which the information was directly presented to the child) and done for about 20% of the instructional period, but unscripted and with the use of leveled texts instead of decodable texts. No major differences in reading outcomes for first graders at risk for reading difficulties were apparent when these two comprehensive programs were compared.

These examples show that the explicitness of instruction may be more important than systematic, scripted lessons in accounting for the effect of systematic phonics. Creating a scope and sequence, using decodable text, and engaging in other ways of systematizing instruction make instruction explicit, but explicitness can be achieved in other ways. Where a teacher operates on the instructional continuum may depend on factors like preparation, experience, the base rate of struggling readers, the school context, and related factors. However, teachers need to be intentionally clear about how the alphabet relates conventionally to sound segments in speech. The supporting materials that are used may vary depending on teacher and student knowledge and skills.

In contrast to the seemingly endless political and ideological commentaries about the purposes and findings of the NRP study, Camilli et al. (2003, 2006) and Hammill and Swanson (2006) have advanced the field by focusing on the data and the need for replication and confirmation of the NRP findings. Adjudication of the issues raised by the NRP report through attempts at replication and continued experimentation can help move the field beyond the simplistic instructional dichotomies that have plagued theory and instruction on reading toward richer and more complex approaches that will enhance reading proficiency for all children. Nonetheless, our analyses of Camilli et al. (2003, 2006) and Hammill and Swanson do not support the belief that the NRP misrepresented their findings or misled policymakers and the educational community (Allington, 2006). The NRP relied on empirical synthesis (meta-analytic methods) for the interpretation of a large body of research. The NRP report explicitly stated the criteria for including studies in the meta-analysis and was subjected to peer review prior to its release to the public.<sup>2</sup> Our reanalysis of the NRP findings confirm their conclusions concerning phonics instruction but must be understood in the context of the need for comprehensive approaches to reading instruction. As the NRP (NICHD, 2000, p. 2-97) stated, "Phonics instruction is never a total reading program," and it "should be integrated with other reading instruction."

These conclusions lead to what we believe should be the reading community's vision of an effective reading program. That is, comprehensive instruction involves explicit instruction in the al-

phabetic principle, explicit instruction in comprehension and vocabulary, and active engagement of the child to develop fluency (Pressley, 2005; Snow et al., 1998). Few in the reading community would disagree that this task is arduous and hardly begun, and the next step is advancing beyond the findings of the NRP and other consensus reports. When phonics is systematic (as defined by the NRP), additional well-conceived literacy activities (as defined by Camilli et al., 2003, 2006) are added, and tutoring is used to increase intensity, the effect sizes may be larger than for any of these components in isolation. That is the important message of the NRP report, Camilli et al. (2003, 2006), and Hammill and Swanson (2006). Although it seems difficult to move beyond the historic dichotomy of reading instructional approaches, it is time to embrace comprehensive approaches to reading instruction and work toward determining how to integrate different components of reading instruction into classroom practice so that the diversity of students and their individual needs can be addressed.

<sup>2</sup> Camilli et al. (2006, p. 30) were in error when they indicated that the NRP report was not subjected to peer review prior to its release (P. McCordle, personal communication, February 21, 2007). In suggesting that the NICHD change procedures for producing meta-analyses, there is a misunderstanding. Consensus reports at the National Institutes of Health are usually done by the Office of Medical Applications of Research (OMAR), which convenes panels of scientists to produce consensus reports. Although the NRP was congressionally mandated, it preceded OMAR in deciding to use meta-analysis. Procedures for conducting syntheses, including the use of meta-analysis, are determined by the specific committee, not the National Institutes of Health or OMAR.

## References

- Allington, R. L. (2002). *Big brother and the national reading curriculum: How ideology trumped evidence*. Portsmouth, NH: Heinemann.
- Allington, R. L. (2006). Reading lessons and federal policy making: An overview and introduction to the special issue. *Elementary School Journal*, 107, 3-15.
- American Federation of Teachers. (1999, June). *Teaching reading is rocket science: What expert teachers of reading should know and be able to do*. Washington, DC: Author.
- Camilli, G., Vargas, S., & Yurecko, M. (2003). Teaching children to read: The fragile link between science and federal education policy. *Education Policy Analysis Archive*, 11(15). Retrieved March 20, 2007, from <http://epaa.asu.edu/epaa/v11n15/>
- Camilli, G., Wolfe, P. M., & Smith, M. L. (2006). Meta-analysis and reading policy: Perspectives on teaching children to read. *Elementary School Journal*, 107, 27-36.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Connor, C. M., Morrison, F. J., Fishman, B. J., Schatschneider, C., & Underwood, P. (2007). Algorithm-guided individualized reading instruction. *Science*, 315, 464-465.
- Cooper, H. (2005). Reading between the lines: Observations on the report of the National Reading Panel and its critics. *Phi Delta Kappan*, 86, 456-461.
- Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research*, 71, 393-447.
- Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps chil-



- dren learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly*, 36, 250–287.
- Elbaum, B., Vaughn, S., Hughes, M. T., & Moody, S. W. (2000). How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research. *Journal of Educational Psychology*, 92, 605–619.
- Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology*, 90, 37–55.
- Foorman, B. R., Schatschneider, C., Eakin, M. N., Fletcher, J. M., Moats, L. C., & Francis, D. J. (2006). The impact of instructional practices in Grades 1 and 2 on reading and spelling achievement in high poverty schools. *Contemporary Educational Psychology*, 31, 1–29.
- Garan, E. M. (2001). Beyond the smoke and mirrors. *Phi Delta Kappan*, 82, 500–506.
- Hammill, D. D., & Swanson, H. L. (2006). The National Reading Panel's meta-analysis of phonics instruction: Another point of view. *Elementary School Journal*, 107, 17–26.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Laird, J., DeBell, M., & Chapman, C. (2006). *Dropout rates in the United States: 2004* (NCES 2007–024). Retrieved March 20, 2007, from <http://nces.ed.gov/pubsearch>
- Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist*, 48, 1181–1209.
- Mathes, P. G., Denton, C. A., Fletcher, J. M., Anthony, J. L., Francis, D. J., & Schatschneider, C. (2005). An evaluation of two reading interventions derived from diverse models. *Reading Research Quarterly*, 40, 148–183.
- National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00–4769). Washington, DC: U.S. Government Printing Office.
- Pedhazur, E. J. (1982). *Multiple regression in behavioral research* (2nd ed.). Fort Worth, TX: Holt, Rinehart, & Winston.
- Prentice, D., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, 112, 160–164.
- Pressley, M. (2005). *Reading instruction that works: The case for balanced instruction* (3rd ed.). New York: Guilford Press.
- RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND.
- Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest*, 2, 31–74.
- Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, 74, 166–169.
- Shanahan, T. (2004). Critiques of the National Reading Panel report: Their implications for research, policy, and practice. In P. McCardle & V. Chhabra (Eds.), *The voice of evidence in reading research* (pp. 235–265). Baltimore: Brookes.
- Shaywitz, S. E. (2004). *Overcoming dyslexia*. New York: Knopf.
- Snow, C. E., Burns, S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.
- Stanovich, K. E. (2000). *Progress in understanding reading: Scientific foundations and new frontiers*. New York: Guilford Press.
- Steering Committee of the Physicians' Health Study Research Group. (1989). Final report on the aspirin component of the ongoing Physicians' Health Study. *New England Journal of Medicine*, 321, 129–135.
- Swanson, H. L., Hoskyn, M., & Lee, C. (1999). *Interventions for students with learning disabilities: A meta-analysis of treatment outcome*. New York: Guilford Press.
- Thorndike, R. (1997). *Measurement and evaluation in psychology and education* (6th ed.). Upper Saddle River, NJ: Prentice Hall.
- Torgesen, J. K. (2000). Individual responses in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research and Practice*, 15, 55–64.
- Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K. S., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities*, 34, 33–58.
- Trusty, J., Thompson, B., & Petrocelli, J. V. (2004). Practical guide for reporting effect size in quantitative research in the *Journal of Counseling & Development*. *Journal of Counseling & Development*, 82, 107–110.
- Vacha-Haase, T., & Thompson, B. (2004). How to estimate and interpret various effect sizes. *Journal of Counseling Psychology*, 51, 473–481.
- Vickery, K., Reynolds, V., & Cochran, S. (1987). Multisensory teaching approach for reading, spelling, and handwriting, Orton-Gillingham based curriculum, in a public school setting. *Annals of Dyslexia*, 37, 189–200.
- Wise, B., Ring, J., & Olson, R. K. (2000). Individual differences in gains from computer-assisted remedial reading with more emphasis on phonological analysis or accurate reading in context. *Journal of Experimental Child Psychology*, 77, 197–235.

Received April 16, 2007

Revision received August 11, 2007

Accepted August 30, 2007 ■

# Text Comprehension in Chinese Children: Relative Contribution of Verbal Working Memory, Pseudoword Reading, Rapid Automatized Naming, and Onset-Rime Phonological Segmentation

Che Kan Leong

University of Saskatchewan and Chinese University of  
Hong Kong

Shek Kam Tse and Ka Yee Loh

University of Hong Kong

Kit Tai Hau

Chinese University of Hong Kong

The present study examined the role of verbal working memory (memory span, tongue twister), 2-character Chinese pseudoword reading, rapid automatized naming (letters, numbers), and phonological segmentation (deletion of rimes and onsets) in inferential text comprehension in Chinese in 518 Chinese children in Hong Kong in Grades 3 to 5. It was hypothesized that verbal working memory, together with a small contribution from the other constructs, would explain individual variation in the children's text comprehension. Structural equation modeling and hierarchical multiple regression analyses generally upheld the hypotheses. Though Chinese pseudoword reading did not play an important mediating role in the effect of verbal working memory on text comprehension, verbal working memory had strong effects on pseudoword reading and text comprehension. The findings on the Chinese language support current Western literature as well as display the differential role of the constructs in Chinese reading comprehension.

*Keywords:* Chinese text comprehension, verbal working memory, pseudoword reading

Children's reading comprehension has been shown to be influenced by their verbal working memory (Cain, Oakhill, & Bryant, 2000, 2004); their rapid, automatic decoding of words (Perfetti, 1985); their phonological skills (Shankweiler, 1989); and a range of cognitive and metacognitive strategies, such as monitoring comprehension, activating background information, integrating multiple strategies of questioning, clarifying and searching for information, identifying story themes, summarizing main points, and predicting outcomes (e.g., Oakhill, Cain, & Bryant, 2003; Perfetti, Landi, & Oakhill, 2005; Williams et al., 2002).

Che Kan Leong, Department of Educational Psychology and Special Education, University of Saskatchewan, Saskatoon, Saskatchewan, Canada, and Department of Educational Psychology, Chinese University of Hong Kong, Shatin, Hong Kong, China; Shek Kam Tse and Ka Yee Loh, Faculty of Education, University of Hong Kong, Hong Kong, China; Kit Tai Hau, Department of Educational Psychology, Chinese University of Hong Kong, Shatin, Hong Kong, China.

We thank the children, their parents, and their teachers in the two schools in Hong Kong for their participation in the project, and we thank the research assistants for their work with the children. Che Kan Leong is grateful to the Social Sciences and Humanities Research Council of Canada for its support through Research Grant 410-01-0059. Shek Kam Tse, Ka Yee Loh, and Kit Tai Hau acknowledge the assistance of internal university grants.

Correspondence concerning this article should be addressed to Che Kan Leong, Department of Educational Psychology and Special Education, University of Saskatchewan, 28 Campus Drive, Saskatoon, Saskatchewan S7N 0X1, Canada. E-mail: chekan.leong@usask.ca

On Chinese children's reading acquisition and development in Chinese, almost all the studies are on phonological sensitivity, with some data on orthographic and morphological processing in relation to reading characters or two-character words (e.g., Ho & Bryant, 1997; McBride-Chang et al., 2005; McBride-Chang & Ho, 2005; Shu, Chen, Anderson, Wu, & Xuan, 2003; Siok & Fletcher, 2001). There are few experimental studies on reading comprehension in Chinese elementary school students. Drawing on the various studies with children using alphabetic language systems as background, we examined the relative contribution of verbal working memory, two-character Chinese pseudoword reading, rapid naming, and phonological sensitivity skills to Chinese text comprehension in a sample of 518 elementary school Chinese students in Hong Kong. We hypothesized that Chinese children's text comprehension in Chinese, as assessed by their short, written answers to open-ended inferential questions on literal aspects, coherent links of text, and application of information in a novel situation, would be primarily influenced at the lower level by verbal working memory. We also examined whether the effect of verbal working memory on text comprehension was mediated by Chinese pseudoword reading (Holmbeck, 1997). To a much smaller extent, rapid naming and onset-rime phonological segmentation tasks might also play a role. The appropriate statistical techniques used to examine the relative contributions of the cognitive and linguistic tasks were a set of hierarchical multiple regression analyses. To test mediated effects, we used structural equation modeling, as recommended by Holmbeck (1997).



## Comprehending Chinese Text in Relation to Cognitive and Linguistic Skills

The emphasis of this report is on text inferencing and on some lower level cognitive and linguistic skills that might influence higher level text comprehension. Inference was defined operationally as encoding the information activated during reading, which may not be explicitly stated in the text (Kintsch, 1994; McKoon & Ratcliff, 1989; see also Singer, 1994; van den Broek, 1994). In Hong Kong, what standardized language achievement tests exist are short, omnibus tests with only a small component on reading comprehension. Thus, text comprehension tests needed to be specially devised on the basis of current theory of comprehension, with a certain amount of face validity to ensure acceptance by teachers, students, and parents (Kintsch & Kintsch, 2005). This latter aspect is particularly important, in that a recent study showed that participants could answer the questions for a well-established standardized comprehension test with above-chance accuracy without reading the passages (Keenan & Betjemann, 2006). In the present study, attempts were made to take into account the findings by Kintsch and Kintsch, Keenan and Betjemann (2006), and others. The logic of the text comprehension task is outlined below, while the details are provided in the *Tasks and Procedure* and *Results* sections.

A total of eight short text passages were specially devised. Each text passage was followed by three kinds of questions: one each on literal inference, coherence inference, and elaboration inference. Literal inferencing is concerned with the surface level of comprehension. Coherence inferencing relates to the whole text and is a function of the local, cohesive nature of individual events, which, in turn, involves linguistic devices such as anaphoras, intersentential connectives, and other discourse pointers (Halliday & Hasan, 1976; Sanders, Spooren, & Noordman, 1992). Elaborative inferencing refers not only to information in the propositions or concepts but also to probable expectations about subsequent information that is inferred from the text and knowledge of the comprehender (Kintsch, 1994; Noordman & Vonk, 1999). Examples of these kinds and levels of inferencing are further discussed in the *Tasks and Procedure* section. In the next several sections, we discuss the role of verbal working memory, word reading, rapid naming, and phonological segmentation in text comprehension.

### Verbal Working Memory

*Working memory* refers to processing resources of limited capacity that individuals need to maintain information while simultaneously acting on the same or other information, and it requires the coordination of both storage of information and processing of additional and sometimes unrelated information (Baddeley, 1986). Verbal working memory tasks generally require children to hold increasingly complex verbal information in memory while responding to questions about the tasks and thus play a role in activating information from long-term memory and in making inferences in text comprehension. Moreover, phonological processing information at the word or subword level has an effect on working memory storage and, indirectly, on reading comprehension.

Following from the early, influential work of Daneman and Carpenter (1980), a number of studies have further shown that a

measure of reading span makes a unique contribution to measures of reading comprehension (e.g., Cain et al., 2000; Cain, Oakhill, & Bryant, 2004; Daneman & Carpenter, 1983; Seigneuric & Ehrlich, 2005; Seigneuric, Ehrlich, Oakhill, & Yuill, 2000). The general idea is that participants have to understand the meaning of each of a group of unrelated sentences to be able to answer a comprehension question (the processing aspect) and, at the same time, to recall the last words in the sentences (the storage aspect). Skilled comprehenders may allocate more working memory resources to text comprehension than to word recognition as compared with less skilled readers (Swanson & Berninger, 1995). Given similar levels of background knowledge, good comprehenders tend to make more integrative inferences than poor comprehenders, who are constrained by working memory to build mental models of text (Oakhill, Cain, & Yuill, 1998). Further, poor comprehenders show specific difficulties in the verbal domain of working memory, which, in turn, relate to impairment in semantic and syntactic skills (Nation, Adams, Bowyer-Crane, & Snowling, 1999). In a meta-analysis involving 77 studies and over 6,000 participants, Daneman and Merikle (1996) reaffirmed the original Daneman and Carpenter (1980) study's finding that verbal process plus storage measures of working memory capacity are good predictors of language comprehension. The process plus the storage span measures were used in the present study.

### Chinese Pseudowords

The notion and characteristics of Chinese pseudowords are quite different from those in English. Pseudowords in English are pronounceable nonwords (e.g., *bave*) and have been shown to correlate highly with real word recognition and reading comprehension (e.g., Rack, Snowling, & Olson, 1992). Children's level of skill in reading pseudowords is an indication of their phonological processing ability, which is critical in reading alphabetic orthographies. Does similar logic apply to reading pseudowords in Chinese?

In the first place, a distinction must be made between a Chinese character (*zi*) as a basic orthographic unit or graphic symbol and a word (*ci*), consisting usually of two or more characters. The basis of compositionality of characters is the corpus of about 560 foundational *bujians*, which subsume about 212 *radicals*, with constituent phonetic and semantic cues as aids in reading and spelling and in accessing dictionaries (Chinese National Language Committee, 1998). A character almost always corresponds to a morpheme (not a phoneme) in the spoken language, whereas a word is the smallest independent unit of meaning and is polymorphemic (Leong, 1997). For example, the word *library* consists of three characters meaning *picture*, *book*, and *institution*.

In the present study and in a study by Leong, Cheng, and Tan (2005), two-character Chinese pseudowords, or, strictly, *pseudocis*, were used. For these two-character pseudowords, each of the constituent characters was a real, pronounceable character, but their combination yielded a pronounceable but meaningless Chinese word. This method of constructing Chinese pseudowords gets around the issue of pronounceability, since each constituent character is pronounceable. More important, unlike pseudowords in English, the focus is on the children's correct oral reading of each of the two characters singly and in combination by utilizing



phonetic, orthographic, morphological, and meaning cues. The mechanism of retrieving all linguistic cues from long-term memory to address the phonology in pronouncing the two-character pseudowords may approximate the mechanism in retrieving, integrating, and interpreting information in short Chinese text materials.

### *Rapid Automatized Naming (RAN)*

The early notion by Denckla and Rudel (1974) of “lack of automaticity” as a correlate of reading and its difficulties and the resultant RAN test with practical applications have been substantiated and refined in a number of recent studies. While there have been different studies and debates on the continuous format (presenting all items) or discrete format (presenting items individually) of task presentation (Denckla & Cutting, 1999; van den Bos, Zijlstra, & Spelberg, 2002) and different arrangements of visual arrays of the items (Compton, Olson, DeFries, & Pennington, 2002), it has been shown consistently that RAN discriminates between good and poor readers and adds uniquely to normal elementary reading beyond phonological sensitivity and memory tasks.

RAN, with its presymbolic component (colors and objects) and symbolic component (numbers and letters), is part of an efficient and rapid phonological retrieval process, and the underlying process relating RAN performance to word reading is a complex one (Compton et al., 2002; Denckla & Cutting, 1999; van den Bos et al., 2002). A recent neuroimaging study of 12 adults showed that RAN taps into the same neural network required for complex reading tasks and that the RAN letter task has greater predictive ability because it is sensitive to the posterior areas associated with the reading network (Misra, Katzir, Wolf, & Poldrack, 2004).

### *Onset–Rime Phonological Segmentation Tasks*

On the question of the involvement of phonology in reading Chinese, it is often assumed (erroneously) that reading Chinese characters relies primarily on “visual” skills and orthographic analysis and that the processing route is more from graphic symbols to meaning. While the square-shaped Chinese characters, which occupy the same geometric space for each symbol, are visually complex as compared with English words, there is little support for the assertion that the identification of characters is from graphic symbols directly to meaning. Research using different experimental paradigms, such as priming and backward and forward masking, has shown that phonology is an interactive constituent part in identifying Chinese characters and is activated early, rapidly, and at the moment of recognizing orthographic shapes (“at lexicality”; Spinks, Liu, Perfetti, & Tan, 2000; Tan & Perfetti, 1998, 1999). Furthermore, according to the universal writing system constraint, all writing systems encode language, and according to the universal phonological principle, the activation of word pronunciation occurs across all writing systems. In addition, the effect of phonology is robust as tested with event-related potentials and functional MRI (Perfetti, 2003; Perfetti & Liu, 2005; Perfetti, Liu, & Tan, 2002).

In Chinese, there is evidence that onset–rime deletion, rather than phonemic awareness, of Chinese characters predicts Chinese

character and word reading (Siok & Fletcher, 2001). This emphasis on onset–rime deletion, rather than segmental phoneme deletion, is in keeping with psycholinguistic analysis that Chinese is basically paradigmatic and not so much segmental and that the basic unit of the character is coterminous with the syllable with its onsets and rimes (Leong, 1997; Wang, 1985). Do these results of the role of onset–rime deletion go beyond character and word reading? Do they apply to reading comprehension in Chinese? Quite possibly the incorporation of phonological segmentation tasks at the beginning phase in reading in Chinese can explain some additional individual variations in reading comprehension in Chinese (McBride-Chang, Bialystok, Chong, & Li, 2004).

## Method

### *Participants*

A total of 518 children in Grades 3, 4, and 5 from two elementary schools in Hong Kong, who were rated as average relative to the territory-wide population in students’ academic performance and their family’s social economic status, took part in the study. In brief, Hong Kong Grade 6 students are broadly classified into three bands according to their territory-wide academic examinations performance, with Band 1 being the highest, Band 2 average, and Band 3 the lowest. The students in the two schools were mostly in Band 2 and, hence, considered to be average in academic performance. In terms of socioeconomic status, the families of these two schools could be rated as at the median household income level according to the 2006 Hong Kong population by census (Hong Kong Census and Statistics Department, 2006). The schools and their students were thus selected on the basis of average academic performance and median socioeconomic status territory-wide, and the students all consented to take part in the study as volunteers.

There were 140 Grade 3 students from six classes ( $n_{\text{boy}} = 82$ ,  $n_{\text{girl}} = 58$ ;  $M_{\text{age}} = 9.06$  years,  $SD_{\text{age}} = 0.44$  years), 187 Grade 4 students from seven classes ( $n_{\text{boy}} = 110$ ,  $n_{\text{girl}} = 77$ ;  $M_{\text{age}} = 10.08$  years,  $SD_{\text{age}} = 0.47$  years), and 191 Grade 5 students from nine classes ( $n_{\text{boy}} = 108$ ,  $n_{\text{girl}} = 83$ ;  $M_{\text{age}} = 11.21$  years,  $SD_{\text{age}} = 0.69$  years). For the total sample, the mean age was 10.22 years ( $SD = 1.02$  years). The group tasks were administered in the classrooms of the students, and the individual tasks were administered in quiet rooms in the school, all in random order. Three full-time and six part-time Chinese-speaking senior research assistants from a local university were given several days’ intensive training on the rationale of the project, the reasons and designs of the group and individual tasks, and specifics of administration before their field work in the schools. These assistants all had experience in working on other projects and, during the intensive interactive training sessions, also offered considerable advice on the tasks, which were further modified and fine-tuned before their administration. These experienced assistants were carefully supervised by Shek Kam Tse and Ka Yee Loh to ensure high fidelity of the field work.

### *Tasks and Procedure*

*Text comprehension.* Because there is no standardized reading comprehension tests for Hong Kong Chinese children, this task had to be specially designed and its psychometric properties as-



certained. Eight short text passages were adapted, modified, and rewritten in traditional (complex) Chinese characters from the most recent and currently used series of Chinese text books for Grades 3 to 6 published by a leading textbook publisher in Beijing. These books were selected as the basis for the text comprehension passages because the original sources were approved by the national textbooks committee of China and the characters and words used were based on statistical analyses of the corpus (Beijing Language Institute, 1984).

Of the eight rewritten text passages, four were narrative pieces (Passages 1, 2, 3, and 8, titled "Pearl of the Orient [Hong Kong]," "Spear and Shield," "Shutting the Pen After Losing the Goat," and "Moonlight Sonata," respectively); three were expository (Passages 4, 6, and 7, titled "The Great Wall [of China]," "Peanuts," and "Nobel," respectively); and one was a poem from the well-known poet Li Po (Passage 5, titled "[Seeing Friend Off]"). These eight short pieces, ranging in length from 6 to 13 sentences, were carefully balanced in syntactic complexity in accordance with general principles in applied linguistics in Chinese (Chao, 1968; Wang, 1985). The contents were familiar to 9- to 11-year-old Chinese children to minimize the impact of background knowledge. The text comprehension task, with the eight passages followed by three open-ended inferencing questions each, was administered as a written task to the whole class of students in 40 min, plus about 8–10 min for two short practice examples to explain clearly the aim of drawing inferences from text. The children were told to read silently each printed text passage, to concentrate on making inferences, and not to worry about sentence construction and spelling. Differential credits of 0, 1, 2, and 3 were awarded according to the implausibility, shallowness, or depth of the short written answer to each question. The maximum score for the whole task was 72 (8 passages  $\times$  9 marks maximum for three questions per passage).

The principles of scoring the written answers on the basis of transforming knowledge and not merely retelling it (Bereiter & Scardamalia, 1987), of explanatory and not just descriptive answers, and of "envisionment" of text worlds (Langer, 1986) characterized the approach to assessing text comprehension in the present study. Passage 2, "Spear and Shield," is shown in the Appendix with its English translation. The first question, on the different uses of the spear and the shield, was an example of literal inferencing; to secure a score of 3, students had to explain briefly the offensive and defensive uses of the two kinds of weapons. To answer the elaboration question, on the reason for the attack tank being a powerful weapon, students had to integrate the different concepts to go beyond the information given to attain the top score of 3. The third question, about giving a suitable title and giving reasons for the choice, was an example of coherence inferencing, and students had to draw on cohesive concepts in the passage as a whole to explain how the ideas of the spear and the shield could evolve into the powerful attack tank. Another concrete example came from Li Po's poem on seeing a friend off; the first question, on the direction of the house where the farewell was said, was given in the first line and was an example of literal inferencing. The question on the imagery of the firework exemplified coherence inferencing. The question on the poet's feeling in bidding a good friend farewell drew on the emotive aspect and required elaboration and integration of different concepts conveyed explic-

itly and implicitly in the short poem. To score the maximum of 3 on the question on the firework, participants had to recognize and explain that the line "Firework haze in Yangzhou in the third moon" suggests spring, fog in March, and the prosperity of Yangzhou.

To ensure consistency of scoring, two research assistants independently scored each set of written protocols according to the marking rubrics explained above. Internal consistency, as measured by Cronbach's alpha, for the eight passages was .908, indicating that these passages as a whole and the answers to the comprehension questions were consistent and reliable. On the basis of the genre and structure, the mainly narrative pieces (Passages 1, 2, 3, and 8) constituted Text Comprehension 1. The interrater reliabilities for the answers to the 12 questions of these four passages were, respectively, .854, .876, .850, and .709. The mainly expository pieces (Passages 4, 5, 6, and 7) formed the second indicator, Text Comprehension 2. The interrater reliabilities for the 12 answers to the questions of these four passages were, respectively, .702, .792, .716, and .657.

To establish some face validity of the text comprehension passages and to ensure that participants could not arrive at the written answers without reading the passages (see Keenan & Betjemann, 2006), we asked a randomly selected group of 53 Grade 4 students ( $M_{\text{age}} = 9.00$  years,  $SD_{\text{age}} = 0.24$  years) who were not in the study simply to write down their answers to the same comprehension questions without having read the passages. A further group of 33 university students ( $M_{\text{age}} = 22.76$  years,  $SD_{\text{age}} = 2.23$  years) was likewise asked to answer the questions without the benefit of reading the passages. These 53 sets of just answers from the Grade 4 children and the 33 sets of just answers from the university students, which they gave without reading the passages, were all scored by the same assistants who worked on the total sample of 518 students. The results of the passage-dependent reading comprehension tasks from these two different groups are discussed in the Results section.

*Memory span task.* The memory span task as a main indicator of verbal working memory for children was modeled after the sentence span task of Swanson (1992), which derives its rationale and format from the influential Daneman and Carpenter (1980) work. A total of 13 sets of two, three, four, and five sentences, all unrelated in meaning, were read orally by the experimenter to the participants as a group task. Participants were asked to listen to each set of sentences and to write down their short answers to a comprehension question and, at the same time, the last word in each sentence of the set. A translated example from a two-sentence set is "The sun gives out bright light. I helped mom do a hard job." The expected answer to the comprehension question "What kind of light does the sun give out?" should be "bright," and the last words should be "light, job." The total testing time for this task was about 25 min, which was found to be ample, and all the answers were scored independently by two assistants. One mark was awarded for each correct answer to each comprehension question and each word correctly named, with a maximum score of 60 (13 answers and 47 last words). The internal consistency Cronbach's alpha of the total answers was .930, while the interrater reliability was .904. A further example of the task is provided in the Appendix.

*Tongue-twister task.* The Chinese tongue-twister task, which is a popular game for both young and old Chinese alike, is based



on the logic that automatic phonemic interference arises more from working memory processes and much less from articulatory processes (Zhang & Perfetti, 1993), and this finding makes it a suitable task to subserve the construct of verbal working memory. Chinese tongue twisters have the added advantage that there is no confounding between phonological and visual–orthographic similarity, and this should produce a stronger tongue-twister effect than in English. Zhang and Perfetti (1993) reported a robust tongue-twister effect for visually presented tongue-twister short stories in Chinese, compared with control stories. These results upheld the phonemic nature of the tongue-twister effect and the activation of the phonological code to support reading comprehension in Chinese.

The present tongue twisters were designed with sets of nonsegmental phonemes, such as alveolar fricatives (/s/ and /z/), alveolar stops (/t/ and /d/), and bilabial and velar stops (/b/ and /p/, /g/ and /k/). The spoken sentences were modifications of those used by Leong and Tan (2002) with Putonghua-speaking children in Beijing, later modified to accommodate Cantonese speech sounds used by children in Hong Kong (see Leong et al., 2005), and were further refined for the present study.

There were eight sets of sentences drawing on discriminating Cantonese speech sounds with their different lexical tones. An example of a two-sentence set of tongue twisters (with their Cantonese lexical tones indicated in square brackets) is “Silzi[2]saanlsoeng[5]silasnlzi[2], saanlzi[2]mun[4]cin[4]-sei[3]silzi[2]saanlzi[2], Sannlzi[2]si[6]sim[4]zi[2], silzi[2]-si[6]sek[6]silzi[2].” The translation is as follows: “Lion Temple is on top of Lion Hill; there are four lion [statues] in front of the temple. The temple is a Buddhist temple; lion statues are stone lions.” The general idea in this actual item was to play on the Cantonese onsets of /s/ and /ts/. A further example is shown in the Appendix. Each child was asked to listen to each spoken sentence and to repeat it verbatim at his or her own pace in the same word order and the same lexical tone as spoken. The tongue twisters were scored according to the number of characters repeated in the correct order and tone, and the maximum was 170. Testing time was about 15 min. Internal consistency Cronbach’s alpha was .808.

*Chinese pseudoword reading.* The Chinese pseudoword reading task consisted of 72 items (sample items appear in Appendix), with the characters all carefully selected from the same series of textbooks, published by People’s Education Publishing in Beijing, on which the comprehension passages were based. Each of the two constituent characters was a real character, but their combination yielded a pronounceable, meaningless Chinese pseudoword. The 72 two-character pseudowords task was refined from that used by Leong et al. (2005) with 157 Chinese students in Grades 4 and 5 in Beijing and Hong Kong, which, in turn, was derived from the 36 most discriminating items from a 72-item task for Grade 4 children and another 36 most discriminating items from a 72-item task for Grade 5 children. The refined 36 items for Grade 4 constituted subtask Pseudoword 1 (PW1), and the refined 36 items for Grade 5 formed subtask Pseudoword 2 (PW2). PW1 and PW2 also took into account such linguistic principles as printed frequency and lexical tones of the characters (Chao, 1968). Each child was asked to read aloud correctly and rapidly each decontextualized two-character pseudoword or pseudo-*ci*. Total testing time was about 7 min per child. A credit of 1 was given for each character identified

and read correctly, and the maximum score for both PW1 and PW2 was 144. Cronbach’s alpha coefficient for the total task was .937.

*RAN.* From the studies discussed earlier and the factor analytic finding by van den Bos et al. (2002) of the separability of presymbolic and symbolic components (see also van den Bos, Zijlstra, & van den Broeck, 2003), the alphanumeric part of RAN (one for letters of the alphabet, and one for Arabic numbers) was administered individually to the participants. The format used was the alternative version (RAN alternative) of Compton et al. (2002), because of their finding that this arrangement explained significantly more variance in word recognition and orthographic processing skills, as compared with the traditional arrangement.

Following this logic and using the same items and arrangement from Compton et al. (2002), our RAN letter naming (RANL) task consisted of the high-frequency lowercase letters (*a, b, d, o, p, s*), and our RAN number (RANN) task consisted of the six digits (1, 2, 4, 6, 7, 9), all presented in random order in 15 rows of five items each. Letters of the alphabet were selected to preserve the integrity of the original task and because letter names are learned and overlearned by children in Hong Kong from kindergarten onward. Any load on memory would be reduced, as the task could be done almost automatically. The alternative of substituting the familiar letters of the alphabet with Chinese characters might approximate the “alpha” aspect but run the risk of changing the structure of the tasks and of introducing the element of unfamiliarity and the need to draw on long-term memory.

The individual children were asked to read horizontally from left to right across the printed page and to name the numbers and letters as rapidly and as accurately as possible in two separate sessions of 30 s each. The total score correct within this time limit for each component was taken as the RANL or RANN score of the child. The use of scores in correctly naming the numbers or letters in unit time (30 s) made for easier administration and obviated the interpretation of negative scores if time in seconds was used as the metric. The maximum score for each part was 75, and the internal consistency Cronbach’s alpha of the two parts was .738.

*Onset–rime phonological segmentation.* The speech-sound segmentation construct for the present study was subserved by two tasks at the syllabic level: deletion of rime, and deletion of onset (see sample items in the Appendix). For deletion of rime, 10 items were spoken Chinese characters, and 10 were spoken English monosyllabic words known to the children (e.g., /m-ian/, /h-ide/). Similarly, for the onset deletion, there were 10 spoken Chinese characters and 10 spoken English words familiar to the children (e.g., /t-ian/, /g-old/). This task was first devised and used successfully by Leong and Tan (2002) in predicting Chinese word reading in two studies of Chinese children, one with 32 Grade 4 students and 38 Grade 5 students, and the other with 180 students in Grades 3, 4, and 5. The further use of the onset–rime deletion task in its present content and format to examine its role in Chinese text comprehension seemed justified. The maximum score for rime deletion and onset deletion was 20 each. Individual children listened to the spoken character or one-syllable word and were asked to delete the end sound (deletion of rime) or the beginning sound (deletion of onset) and to say what was left. Internal consistency Cronbach’s alpha for the segmentation task as a whole was .766.



## Results

### Preliminary Analyses

Table 1 displays the means and standard deviations of the various measures used in the study. The intercorrelations of these tasks after we controlled for chronological age for the total group of 518 students are shown in Table 2. As all tests were group or individually administered, there were no missing data in the data set. Preliminary scanning showed that all variables were practically normally distributed. In the structural equation analyses, the maximum likelihood fitting function was used, which was demonstrated to be quite robust to data violating normality assumptions even at small sample sizes (Hau & Marsh, 2004).

In a 3 (grade level: Grades 3, 4, 5)  $\times$  2 (gender)  $\times$  8 (text comprehension passage) multivariate analysis of variance with passage as a within-subject variable, Wilks's lambda was significant for grade,  $F(2, 517) = 11.503$ ,  $p < .001$ ,  $\eta^2 = .154$ , and gender,  $F(2, 517) = 5.047$ ,  $p < .001$ ,  $\eta^2 = .074$ . There was no Grade  $\times$  Gender interaction. Further detailed between-subjects analyses by each of the passages showed that the effect sizes for grade ranged from .064 to .217, with a mean of .104, and that the effect sizes for gender ranged from .001 to .030, with a mean of .009. Because of the very low effect sizes for gender (Cohen & Cohen, 1983), subsequent analyses concentrated on the total group of 518 students.

As discussed earlier, one important question raised recently in the reading research literature regards the passage independence of items (whether participants can answer items with above chance accuracy without reading the passages), as shown and discussed by Keenan and Betjemann (2006) and Kintsch and Kinstch (2005). To address this issue and to ensure that the specially designed Chinese text comprehension task, with its eight passages and 24 questions, was passage dependent (i.e., participants needed to read the passages to answer the questions correctly), we found that the 53 Grade 4 children not in the study who just answered the questions without reading the passages obtained a mean score of 5.377 (or 0.07 out of a maximum of 72), with a standard deviation of 2.90. The 33 university students who also just answered the questions without reading the passages obtained a mean of 25.06 (0.35 out of

a maximum of 72) and a standard deviation of 8.047. The overall performance of 0.07 by the Grade 4 students and the less than chance performance by the university students suggest that the passages needed to be read before the open-ended questions could be answered correctly. The significance of this issue is further discussed in the *Summary and Conclusion* section.

As a preliminary analysis on the appropriateness of the categorization of the tasks by their latent factor, a confirmatory factor analysis was conducted on all the tasks used in the study (see Table 3). The latent endogenous construct of text comprehension was subserved by Text Comprehension 1 and Text Comprehension 2. The latent pseudoword reading construct was subserved by PW1 and PW2. For the exogenous constructs, verbal working memory was subserved by the measurable tasks of memory span and tongue twister, RAN was subserved by RANL and RANN, and onset-rime segmentation was subserved by the deletion of rime and deletion of onset.

On the basis of the various goodness-of-fit indexes recommended by Marsh, Hau, and Grayson (2005), the fit of the model to the total group of 518 students was satisfactory,  $\chi^2(25) = 33.21$ ,  $p = .124$  (root-mean-square of approximation [RMSEA] = .025; nonnormed fit index [NNFI] = .996; comparative fit index [CFI] = .998). These fit indexes reflect the appropriateness of the tasks as measurable indicators of the various latent constructs. All the factor loadings of the tasks, with the exception of the slightly lower value for the tongue twister, were at least .70, further supporting the validity of the posited measures.

### Full Model With Structural Equation Analyses

The strength of relations among the constructs in our hypothetical model (see Figure 1) was examined with structural equation modeling using maximum likelihood estimation (LISREL Version 8.72; Jöreskog & Sörbom, 1996–2001). The model tested is summarized schematically in Figure 1. Though this model looks slightly different from that in the earlier confirmatory factor analysis, statistically they are equivalent, because the structural parts of both models were saturated with freely estimated paths linking all possible pairs of factors (MacCallum, Wegener, Uchino, & Fab-

Table 1  
Means and Standard Deviations of All Tasks for the Total Sample

Task	Maximum	Grade 3 ( <i>n</i> = 140)		Grade 4 ( <i>n</i> = 187)		Grade 5 ( <i>n</i> = 191)		Total ( <i>N</i> = 518)	
		<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Text comprehension total	72	27.57	12.32	37.31	14.53	47.39	18.09	38.39	17.29
Text Comprehension 1	36	18.30	7.97	23.85	8.88	28.77	11.15	24.16	10.41
Text Comprehension 2	36	9.27	5.91	13.46	6.90	18.62	8.29	14.23	8.11
Pseudoword Reading 1	72	34.85	12.78	41.10	11.98	47.57	10.86	41.79	12.82
Pseudoword Reading 2	72	38.19	11.32	42.00	11.43	46.55	10.20	42.65	11.44
Memory span	60	25.73	11.11	27.61	12.16	33.66	13.67	29.33	12.91
Tongue twister	170	136.99	21.06	139.89	20.82	141.71	21.23	139.78	21.08
Rapid automatized naming (letter)	75	52.91	11.35	57.91	11.28	60.89	11.23	57.66	11.70
Rapid automatized naming (number)	75	65.69	9.06	69.95	7.56	71.08	7.31	69.21	8.19
Deletion of rime	20	11.99	5.00	13.01	3.83	14.31	4.20	13.21	4.40
Deletion of onset	20	8.52	5.96	7.90	5.38	10.03	5.83	8.85	5.77

Table 2  
Partial Correlations of Tasks After Controls for Age for the Total Sample ( $N = 518$ )

Task	1	2	3	4	5	6	7	8	9	10	11
1. Text comprehension total	—										
2. Text Comprehension 1	.945	—									
3. Text Comprehension 2	.901	.710	—								
4. Pseudoword Reading 1	.562	.540	.494	—							
5. Pseudoword Reading 2	.522	.502	.458	.889	—						
6. Memory span	.579	.527	.550	.476	.449	—					
7. Tongue twister	.425	.398	.387	.368	.350	.357	—				
8. Rapid naming—letter	.413	.420	.333	.402	.361	.370	.212	—			
9. Rapid naming—number	.328	.311	.293	.306	.252	.279	.204	.607	—		
10. Deletion of rime	.349	.331	.314	.323	.317	.294	.246	.319	.242	—	
11. Deletion of onset	.318	.309	.275	.335	.333	.273	.227	.281	.171	.644	—

Note. All partial correlations are significant at  $p < .001$ .

rigar, 1993). Understandably, the fit of this model (in Figure 1) was equally satisfactory,  $\chi^2(25) = 33.21$ ,  $p = .124$  (RMSEA = .025; NNFI = .996; CFI = .998). To further test whether RAN and onset-rime segmentation might have some indirect effect via pseudoword reading on text comprehension, we put both of them in the model for testing. Empirical results suggested that neither of them had such an indirect effect. When all the nonsignificant paths were removed, the fit of the model was still very good, with the major paths practically unchanged,  $\chi^2(30) = 37.48$ ,  $p = .164$  (RMSEA = .022; NNFI = .997; CFI = .998). This gives further credence to our hypothetical model as depicted in Figure 1.

Results showed that verbal working memory, RAN, and onset-rime segmentation were highly and positively correlated at around .4 to .6 (all  $ps < .01$ ; see Figure 1). However, despite their close positive relations, they had disparate effects on text comprehension and pseudoword reading. In general, with the exception of a very small positive effect ( $\beta = .11$ ,  $ns$ ) of RAN on pseudoword reading, RAN and onset-rime segmentation did not have any appreciable effect on text comprehension or pseudoword reading (all paths were nonsignificant).

Verbal working memory, however, had much stronger effects (see Figure 1). Children with better verbal working memory performed much better in pseudoword reading ( $\beta = .62$ ,  $p < .01$ ) and text comprehension ( $\beta = .83$ ,  $p < .01$ ). Though it is speculated that pseudoword reading might play the mediating role in the effect of verbal working memory on text comprehension, there is not much empirical support. If pseudoword reading had been the mediating factor, the path between pseudoword reading and text comprehension would have been much stronger (it was only .06, nonsignificant). Thus, verbal working memory helped children's text comprehension performance quite independently and directly, rather than through improving their pseudoword reading competence, at least for the tasks used in the present study.

#### *Hierarchical Regression Analyses on Text Comprehension Total Score*

Understandably, some of the tasks shared common variance in explaining children's text comprehension. To examine in greater

Table 3  
Confirmatory Factor Analyses of All Tasks in the Study for the Total Sample ( $N = 518$ )

Task	Factor loading				
	Verbal working memory	Rapid automatized naming	Onset-rime segmentation	Pseudoword reading	Text comprehension
Memory span	.719				
Tongue twister	.493				
Rapid automatized naming—letter		.893			
Rapid automatized naming—number		.696			
Deletion of rime			.854		
Deletion of onset			.754		
Pseudoword Reading 1				.988	
Pseudoword Reading 2				.899	
Text Comprehension 1					.877
Text Comprehension 2					.845
Correlations among factors					
Rapid automatized naming	.584				
Onset-rime segmentation	.524	.425			
Pseudoword reading	.710	.494	.419		
Text comprehension	.890	.552	.456	.664	



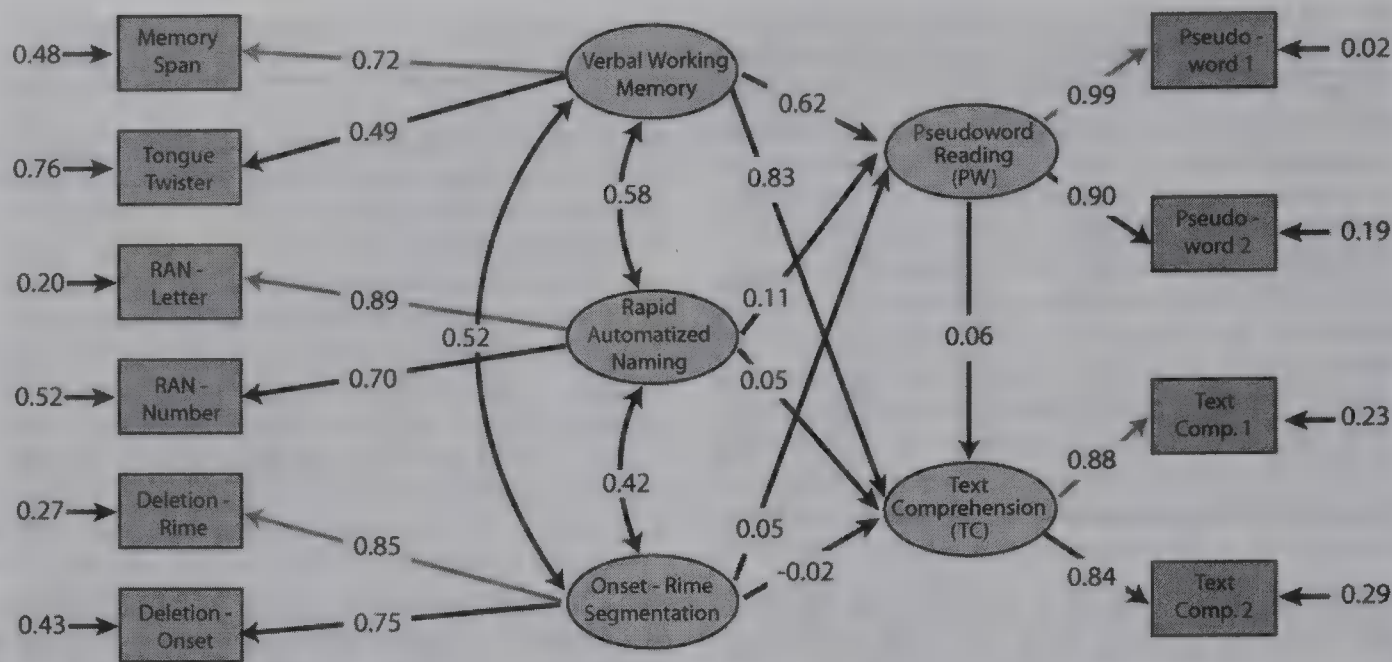


Figure 1. Structural equation modeling showing standardized effects of verbal working memory, rapid automatized naming (RAN), and onset-rime segmentation on text comprehension and pseudoword reading.  $N = 518$ ;  $\chi^2 = 33.21$ ;  $df = 23$ ;  $p = .124$ ; root-mean-square of approximation = .025.

detail the unique and common contributions of each task, we conducted hierarchical multiple regression analyses with text comprehension (sum of standardized scores of the two respective tasks) as the criterion variable. We controlled age by entering it first, while we entered the tasks in verbal working memory, pseudoword reading, RAN, and onset-rime segmentation separately (see Table 4, Models 2, 3, 4, and 5), in pairs (see Table 4, Models 6, 7, 8, 9, 10, 11, and 12), and all together (see Table 4, Model 13).

As can be seen from the analyses (see Table 4, Models 2, 6, 7, 8, 10, and 13), verbal working memory, in particular memory span ( $\beta$ s ranged from .306 to .466), substantially predicted text com-

prehension. Though these working memory tasks shared some of the variance when put jointly with other tasks, working memory's unique contributions to text comprehension remained relatively high. In comparison, PW1 also played a role in text comprehension when considered conjointly with the two working memory tasks (Model 6). It should be noted that after we adjusted for age, squared multiple correlations were .471 for the two memory tasks by themselves (Model 2), .407 for the two pseudoword tasks by themselves (Model 3), and .536 for the four tasks from the two constructs (Model 6).

In contrast, RAN and onset-rime segmentation had much smaller unique contributions in predicting text comprehension.

Table 4  
Hierarchical Multiple Regression Analyses: Predicting Chinese Text Comprehension From Verbal Working Memory, Pseudoword Reading, Rapid Automatized Naming (RAN), and Onset-Rime Segmentation

Model	Age	Predictor ( $\beta$ )								$R^2$
		Memory span	Tongue twister	Pseudoword 1	Pseudoword 2	RAN letter	RAN number	Rime deletion	Onset deletion	
1	.367***									.134
2	.285***	.466***	.231***							.471
3	.209***			.456***	.101					.407
4	.279***					.315***	.119*			.283
5	.336***							.232***	.145**	.248
6	.211***	.341***	.163***	.280***	.041					.536
7	.253***	.399***	.211***			.157***	.059			.501
8	.277***	.425***	.204***					.112**	.062	.491
9	.276***					.241***	.107*	.153**	.112*	.331
10	.252***	.378***	.192***			.129***	.057	.081	.052	.512
11	.187***			.358***	.106	.168**	.076			.446
12	.207***			.417***	.072			.146**	.046	.434
13	.206***	.306***	.148***	.226***	.036	.083*	.046	.070	.019	.550

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

Standing alone (Model 4), RAN, in particular RANL ( $\beta = .315$ ), had some effect on text comprehension. However, its effect dropped moderately to .241 (Model 9) in the presence of onset-rime segmentation tasks and drastically to .157 (Model 7) in the presence of verbal working memory. This shows that the effects of RAN on text comprehension could be accounted for quite substantially by the differences in children's verbal working memory.

Similarly, though onset-rime segmentation, in particular rime deletion ( $\beta = .232$ ; Model 5), seemed to contribute to text comprehension, its effect dropped to .153 (Model 9) with RAN and to .112 (Model 8) with verbal working memory. In the last regression analysis (Model 13), in which all tasks predicted text comprehension together, it was obvious that verbal working memory had the largest unique contribution ( $\beta = .306$  for memory span;  $\beta = .148$  for tongue twister). Though the regression analysis (Model 13) suggested that one of the pseudoword reading tasks might have some unique contribution ( $\beta = .226$  for Task 1), its additional explanatory power was small (the squared multiple correlation increased from 53.6% to 55.0% in Models 6 and 13 when PW1 and PW2 were added). The details are summarized in Table 4.

In sum, the regression analyses with the total text comprehension and its component scores showed that memory span consistently had the largest effect, followed by PW1, on text comprehension. Noticeably, to a large extent, the effects of RANL and rime deletion were shared by the effects of memory span, while the effect memory was quite unique and unexplainable by the other measures. Even though pseudowords tended to have some effects on text comprehension, their additional explanatory power, beyond that of verbal working memory, RAN, and onset-rime segmentation, was relatively small in this particular study on Chinese language.

## Discussion

Inspection of the means and standard deviations of the tasks (see Table 1) shows that the text comprehension mean performance (percentage correct) was at the 53% level for the total group of 518 students, and there was the expected progression from Grades 3 through 4 to 5. One reason is that the passages were deliberately designed to be more stringent and discriminating to avoid ceiling and floor effects, which might arise with the spread of about 2 years in age. The other reason that the students could only get half of the items correct is that elementary Chinese students in Hong Kong are not well guided in their everyday learning of Chinese text materials to analyze meaning construction. They are not well taught to carry out reasoning operations, such as making assumptions, citing and validating evidence, and envisioning the writer's text world (Langer, 1986). Careful inspection of some of the written answers shows that the low-scoring ones were invariably straightforward reproductions or literal translations from the text materials, whereas the high-scoring ones took into account the structure, the gist of the essays, and the intention of the authors.

These are some examples from the written protocols of the students. In the essay on shutting the barn door after the goat had bolted (Passage 3), high scorers went beyond verbatim information to convey the idea of learning from experience, whereas low scorers simply stated the physical act of shutting the barn door. For Passage 4, "The Great Wall of China," an answer to the question

on comparing the Great Wall to a long dragon that brought out the meandering nature of the long wall in the metaphorical splendid dragon, which is also a symbol of strength and prosperity in the Chinese culture, would fetch the top credit of 3. The mere statement that the Great Wall is long and winding would fetch a score of 1. Also from Table 1, the two pseudoword reading tasks were at about a mean of 60% accuracy overall, whereas memory span was at about 49% accuracy overall. In some contrast, the phonological sensitivity tasks and RAN were generally scored higher.

Table 2 shows that the cognate tasks within each latent factor correlated highly to moderately among one another after chronological age was controlled (e.g., .889 for the two pseudoword reading tasks, .357 for the verbal working memory tasks, .607 for the two RAN tasks, and .644 for the two phonological segmentation tasks). The confirmatory factor analysis reaffirms that the indicators of verbal working memory, pseudoword reading, RAN, phonological segmentation, and text comprehension all clustered to form the components as hypothesized, with very high loadings. This lends credence to our conceptualization of the constructs as discussed earlier.

### *Contribution of Verbal Working Memory to Chinese Text Comprehension*

In the literature on reading comprehension in English, there is evidence of the important role of working memory (Cain, Oakhill, & Bryant, 2004; Cain, Oakhill, & Lemmon, 2004; Seigneuric & Ehrlich, 2005; Seigneuric et al., 2000). The storage and manipulation of linguistic materials in memory, as assessed by the kind of memory span task used, likely draw on the same or similar sorts of strategies in processing text materials and represent the resources available for text processing. There is evidence of this in the present study, and the effect of verbal working memory on text comprehension seems to apply across writing systems to both the alphabetic English and the morphosyllabic Chinese.

As shown in Table 4, as the latent constructs and tasks shared and explained jointly Chinese text comprehension, their effects varied somewhat according to what tasks were put together in the analyses. It is important to note that the verbal working memory construct uniquely contributed more to the total effects. These results are in congruence with those of the longitudinal investigation of 74 French children in Grades 1, 2, and 3 by Seigneuric and Ehrlich (2005). These researchers used a working memory task in which the children had to supply missing final words of unrelated sentences read aloud and to recall all the supplied words at the end of the set in the correct order (e.g., "1, 2, 3, 4 are digits," "A, B, C, D are letters," and "To cut meat we use a knife"). What is important is that Grade 2 working memory contributed significantly to Grade 3 reading comprehension. These authors explained that as word recognition became more automatic through the early grades, working memory became a more important determinant in reading comprehension. These results were obtained from a developmental study with 74 French children followed through Grades 1, 2, and 3; the present cross-sectional study, with a much larger number of Chinese students (518 in all), shows very similar patterns.

The structural equation modeling (see Figure 1) provides some answers to the direct and indirect effects of verbal working mem-



ory as a construct on text comprehension. The significant and high path coefficients of .83 from verbal working memory to text comprehension and of .62 from verbal working memory to pseudoword reading suggest the potent effects of the working memory construct on these two endogenous constructs. Variations in levels of the exogenous construct of working memory thus account significantly for the endogenous construct of text comprehension and also pseudoword reading. It is noteworthy that pseudoword reading does play a very small role in Chinese text comprehension (Holmbeck, 1997). It could well be that there are other constructs influencing text comprehension that were not tested here.

The better performance of the older students in inferencing might have been due to their better general knowledge, their better awareness of when and where to draw inferences, and also enhanced processing resources, such as better utilization of verbal working memory. Perusal of sample written scripts shows that students in Grade 3 and also poorer students likely gave verbatim answers taken almost directly from the text passages. They likely failed to draw analogies, such as the analogy shown in the text between the Great Wall of China and the meandering long dragon with all its splendor, prowess, and symbol of prosperity (Passage 4). These younger and poorer comprehenders might also have provided only partial answers, with little or no explanation, when multiple answers with short explanations were called for. These students also failed to grasp the symbolic values of sadness and loss conveyed by such key phrases such as "lone sail" and "distant image [of sailing boat]" in the well-known poem by Li Po, in which he bids farewell to his good friend sailing down the River Yangtze (Passage 5). These students further failed to integrate the local reference to the lowly peanut being buried in the ground and hiding its utility (Passage 6) with the more elaborate text base of being practical and unassuming. In fact, many students simply made reference from their life experience to peanuts used for making peanut butter and for other kinds of food.

### *Chinese Pseudoword Reading*

What are the plausible reasons for Chinese pseudoword reading (PW1) explaining a small amount, jointly with verbal working memory, of the variation in Chinese text comprehension (see Table 4)? At the theoretical level, the interactive constituency model of Tan and Perfetti (1998, 1999) provides the underpinning. This model posits that Chinese word identification results from the convergence of the phonological form, the orthographic form, and the semantic form, with the suggestion that phonologic-orthographic convergence is more rapid and more reliable than the orthographic-semantic. The ability to read accurately and rapidly Chinese pseudowords in the way they were designed for this study likely has some similarity to what is needed to read textual materials with understanding.

The other plausible reason is that to identify and read correctly each character constituting the two-character pseudoword, the child had to draw on his or her knowledge of the vocabulary, which has been shown to affect reading comprehension (Cain et al., 2000; Cain, Oakhill, & Bryant, 2004; Seigneuric & Ehrlich, 2005; Seigneuric et al., 2000). From observations of the performance of the students, these different strategies explained a con-

siderable number of their reading errors: reading by constituent *bujians*, using the first character to infer the second character, reading similarly configured characters, and guessing from certain constituent parts.

Our finding that pseudoword reading made a small contribution to the variation in reading comprehension of Chinese text materials may be compared with the findings regarding reading comprehension in English by Oakhill et al. (2003). These researchers found, in a two-wave study of 102 British children who were 7 to 8 years old at the first wave, a dissociation of word reading and text comprehension, even though these skills were also shown to be correlated. They found that the significant variation in comprehension was explained by tasks of text integration, metacognition monitoring, and working memory (of digits and of sentences and words). It should be noted that the researchers' inference and integration skills tasks were very short (three lines), and the children listened to this short text. In our study, the average length of the eight passages was over eight sentences, and our participants had to read each passage silently and then write their short answers to the inferential questions. Oakhill et al.'s (2003) integration of information and the story anagram tasks emphasized story structure. Our eight passages and 24 open-ended questions were much more challenging intellectually and linguistically. We probed general ideas and asked the children to hypothesize, to use schemas, to cite evidence, and generally to go beyond the information given in making inferences. The issue of the contribution of word reading (not so much vocabulary per se) to differential facets of inferential text comprehension needs further investigation.

### *Role of Phonological Segmentation and RAN*

The phonological sensitivity tasks made almost no contribution to Chinese text comprehension. This pattern likely reflects the much more predominant contributions by verbal working memory, the two-part pseudoword reading tasks, and their conjoint effect. It is also likely that phonological segmentation of onset-rime deletion plays a role only at the emergent literacy stage, as shown by Siok and Fletcher (2001) in their study of preschool and grade school Chinese children in Beijing. More important, our finding that onset-rime segmentation did not affect reading comprehension is in line with the findings with American children by Catts, Fey, Zhang, and Tomblin (1999) and with the findings by Demont and Gombert (1996) in a 4-year follow-up study of 38 French-speaking preschool children.

The low contribution by RAN (both letters and numbers) is in line with general findings from a meta-analysis by Swanson, Trainin, Necoechea, and Hammill (2003) of the correlation literature on measures of reading, RAN, phonological sensitivity, and related abilities from a large study of 2,257 American children in 49 independent samples with corrections for sample size, restriction in range, and attenuations. Swanson et al. found that correlations between RAN and phonological sensitivity were low (.38), and RAN and phonological sensitivity tasks correlated moderately with real word reading (.46 and .48, respectively). RAN and phonological sensitivity were found to be less important than measures of spelling and word attack skills and also played a less important role in reading comprehension. It thus appears that even with quite disparate writing systems—English and Chinese—there



are common and general findings on the role of RAN in reading, if one does not go into the intricacies of different formats of RAN, theoretical underpinnings, and practical applications. In the present study, the two tasks used (RANL and RANN) were easy, were quick to administer (30 s per task per child), and had a role, albeit a small one, in explaining Chinese reading comprehension.

### Summary and Conclusion

To the best of our knowledge, our study with a large sample of 518 Chinese children might be among the first to investigate these children's text comprehension and the relative contribution of some cognitive and linguistic skills. We acknowledge some shortcomings in the study. First, the study was confined to Cantonese-speaking Chinese children in Hong Kong between the ages of 9 and 11 years in Grades 3, 4, and 5 and was not able to assess the developmental paths and possible changes in the children's performance of the different tasks and their interaction. Second, the text-inferencing passages were relatively short, even though the open-ended written answer format was easy to use and showed reasonable content validity. Third, as working memory is a system comprising separable components, multiple tasks and assessments over time are needed to examine alternative theoretical accounts of working memory capacity (Alloway, Pickering, & Gathercole, 2006). Fourth, structural equation modeling, with its composite measurement model and path model (see Figure 1), describes relations of dependency among the latent constructs, and no claim is made to causality. The structural equation model with its acceptable "good" fit and the prediction of Chinese text reading from the different equations of the hierarchical multiple regression analyses (see Table 4) are attempts at answering some research questions and should be interpreted as approximations of reality.

We would, however, also like to suggest that the present study contributes to our knowledge of text comprehension in Chinese in several ways. First, it shows that carefully constructed and relatively short text passages with well-designed, open-ended, inferential questions can tap text comprehension with high fidelity. The related use of the written protocols combining reading and writing was effective in studying text comprehension (see Kintsch & Kintsch, 2005; Swanson & Berninger, 1996). The open-ended inferential questions requiring short written answers showed the passage-dependent nature of the task (Keenan & Betjemann, 2006), as attested by the virtual failure of the random group of 53 Grade 4 students scoring at 7% by just answering the questions. Second, text comprehension has many facets, as shown by our careful study of the written protocols of the students. Third, two-character Chinese pseudowords contributed in some way to reading comprehension, especially in lower elementary grades, thus showing the role of knowing characters and words analytically and synthetically. Fourth, in support of the literature on English reading comprehension, verbal working memory had a strong effect on Chinese text comprehension and, in fact, had the largest joint or unique effects on overall text comprehension. Fifth, again in common with findings from English, there were small, though statistically significant, contributions by rapid naming of letters to the Chinese text comprehension. The challenge is to specify with fine-grained analyses how these various components might relate to different facets of reading comprehension in Chi-

nese and to understand more fully the process of inferencing in text materials as a whole at different elementary grade levels. Moreover, general language mechanisms underlying poor reading comprehension (Nation et al., 1999; Perfetti et al., 2005; Stothard & Hulme, 1992) should be further investigated.

### References

- Alloway, T. P., Pickering, S. J., & Gathercole, S. E. (2006). Verbal and visuospatial short-term and working memory in children: Are they separable? *Child Development*, 77, 1698–1716.
- Baddeley, A. D. (1986). *Working memory*. London: Oxford University Press.
- Beijing Language Institute. (1984). *Hanyu cihui de tongji yu fenxi* [Statistical analyses of Chinese characters]. Beijing, China: Author.
- Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Erlbaum.
- Cain, K., Oakhill, J., & Bryant, P. (2000). Phonological skills and comprehension failure: A test of the phonological deficit hypothesis. *Reading and Writing: An Interdisciplinary Journal*, 13, 31–56.
- Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology*, 96, 31–42.
- Cain, K., Oakhill, J., & Lemmon, K. (2004). Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity. *Journal of Educational Psychology*, 4, 671–681.
- Catts, H. W., Fey, M. E., Zhang, X., & Tomblin, J. B. (1999). Language basis of reading and reading disabilities: Evidence from a longitudinal investigation. *Scientific Studies of Reading*, 3, 331–361.
- Chao, Y. R. (1968). *A grammar of spoken Chinese*. Berkeley: University of California Press.
- Chinese National Language Committee. (1998). *Xun xi chu li yong GB13000.1 zi fu ji: Hanzi bu jian gui fan* [Information processing using GB13000.1 character symbol anthology: Hanzi bujian analyses]. Beijing, China: Beijing Language Institute Press.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Compton, D. L., Olson, R. K., DeFries, J. C., & Pennington, B. F. (2002). Comparing the relationships among two different versions of alphanumeric rapid automatized naming and word level reading skills. *Scientific Studies of Reading*, 6, 343–368.
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19, 450–466.
- Daneman, M., & Carpenter, P. A. (1983). Individual differences in integrating information between and within sentences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 9, 561–584.
- Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, 3, 422–433.
- Demont, E., & Gombert, J. E. (1996). Phonological awareness as a predictor of recoding skills and syntactic awareness as a predictor of comprehension skills. *British Journal of Educational Psychology*, 66, 315–332.
- Denckla, M. B., & Cutting, L. E. (1999). History and significance of rapid automatized naming. *Annals of Dyslexia*, 49, 29–42.
- Denckla, M. B., & Rudel, R. G. (1974). Rapid "automatized" naming of pictured objects, colors, letters, and numbers by normal children. *Cortex*, 10, 186–202.
- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longmans.
- Hau, K. T., & Marsh, H. W. (2004). The use of item parcels in structural



- equation modeling: Nonnormal data and small sample sizes. *British Journal of Mathematical and Statistical Psychology*, 57, 327–351.
- Ho, C. S.-H., & Bryant, P. (1997). Development of phonological awareness of Chinese children in Hong Kong. *Journal of Psycholinguistic Research*, 26, 109–126.
- Holmbeck, G. N. (1997). Toward terminological, conceptual, and statistical clarity in the study of mediators and moderators: Examples from the child-clinical and pediatric psychology literature. *Journal of Consulting and Clinical Psychology*, 65, 599–610.
- Hong Kong Census and Statistics Department. (2006). *2006 population by-census*. Hong Kong: Hong Kong SAR Government.
- Jöreskog, K. G., & Sörbom, D. (1996–2001). *LISREL 8: User's reference guide*. Chicago: Scientific Software International.
- Keenan, J. M., & Betjemann, R. S. (2006). Comprehending the Gray Oral Reading Test without reading it: Why comprehension tests should not include passage-independent items. *Scientific Studies of Reading*, 10, 363–380.
- Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*, 49, 294–303.
- Kintsch, W., & Kintsch, E. (2005). Comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 71–92). Mahwah, NJ: Erlbaum.
- Langer, J. A. (1986). *Children reading and writing: Structures and strategies*. Norwood, NJ: Ablex.
- Leong, C. K. (1997). Paradigmatic analysis of Chinese word reading: Research findings and classroom practices. In C. K. Leong & R. M. Joshi (Eds.), *Cross-language studies of learning to read and spell: Phonologic and orthographic processing* (pp. 379–417). Dordrecht, the Netherlands: Kluwer Academic Publishers.
- Leong, C. K., Cheng, P. W., & Tan, L. H. (2005). The role of sensitivity to rhymes, phonemes and tones in reading English and Chinese pseudowords. *Reading and Writing: An Interdisciplinary Journal*, 18, 1–26.
- Leong, C. K., & Tan, L. H. (2002). Phonological processing in learning to read Chinese: In search of a framework. In E. Hjelmquist & C. von Euler (Eds.), *Dyslexia and literacy* (pp. 126–150). London: Whurr.
- MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, 114, 185–199.
- Marsh, H. W., Hau, K. T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 225–340). Mahwah, NJ: Erlbaum.
- McBride-Chang, C., Bialystok, E., Chong, K. K. Y., & Li, Y. (2004). Levels of phonological awareness in three cultures. *Journal of Experimental Child Psychology*, 89, 93–111.
- McBride-Chang, C., Cho, J.-R., Liu, H., Wagner, R. K., Shu, H., Zhou, A., et al. (2005). Changing models across cultures: Associations of phonological awareness and morphological structure awareness with vocabulary and word recognition in second graders from Beijing, Hong Kong, Korea, and the United States. *Journal of Experimental Child Psychology*, 92, 140–160.
- McBride-Chang, C., & Ho, C. S.-H. (2005). Predictors of beginning reading in Chinese and English: A 2-year longitudinal study of Chinese kindergartners. *Scientific Studies of Reading*, 9, 117–144.
- McKoon, G., & Ratcliff, R. (1989). Semantic associations and elaborative inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 326–338.
- Misra, M., Katzir, T., Wolf, M., & Poldrack, R. A. (2004). Neural systems for rapid automatized naming in skilled readers: Unraveling the RAN–reading relationship. *Scientific Studies of Reading*, 8, 241–256.
- Nation, K., Adams, J. W., Bowyer-Crane, C. A., & Snowling, M. J. (1999). Working memory deficits in poor comprehenders reflect underlying language impairments. *Journal of Experimental Child Psychology*, 73, 139–158.
- Noordman, L., & Vonk, W. (1999). Discourse comprehension. In A. D. Friederici (Ed.), *Language comprehension: A biological perspective* (2nd ed., pp. 229–263). New York: Springer.
- Oakhill, J. V., Cain, K., & Bryant, P. E. (2003). The dissociation of word reading and text comprehension: Evidence from component skills. *Language and Cognitive Processes*, 18, 443–468.
- Oakhill, J. V., Cain, K., & Yuill, N. (1998). Individual differences in children's comprehension skill: Toward an integrated model. In C. Hulme & R. M. Joshi (Eds.), *Reading and spelling: Development and disorders* (pp. 343–367). Mahwah, NJ: Erlbaum.
- Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.
- Perfetti, C. A. (2003). The universal grammar of reading. *Scientific Studies of Reading*, 7, 3–24.
- Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227–247). Oxford, England: Blackwell.
- Perfetti, C. A., & Liu, Y. (2005). Orthography to phonology and meaning: Comparisons across and within writing systems. *Reading and Writing: An Interdisciplinary Journal*, 18, 193–210.
- Perfetti, C. A., Liu, Y., & Tan, L. H. (2002). How the mind meets the brain in reading: A comparative writing systems approach. In H. S. R. Kao, C. K. Leong, & D. G. Gao (Eds.), *Cognitive neuroscience studies of the Chinese language* (pp. 35–60). Hong Kong: Hong Kong University Press.
- Rack, J. P., Snowling, M. J., & Olson, R. K. (1992). The nonword reading deficit in developmental dyslexia: A review. *Reading Research Quarterly*, 27, 29–53.
- Sanders, T. J. M., Spooren, W. P. M., & Noordman, L. G. M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes*, 15, 1–35.
- Seigneuric, A., & Ehrlich, M.-F. (2005). Contribution of working memory capacity to children's reading comprehension: A longitudinal investigation. *Reading and Writing: An Interdisciplinary Journal*, 18, 617–656.
- Seigneuric, A., Ehrlich, M.-F., Oakhill, J. V., & Yuill, N. M. (2000). Working memory resources and children's reading comprehension. *Reading and Writing: An Interdisciplinary Journal*, 13, 81–103.
- Shankweiler, D. (1989). How problems of comprehension are related to difficulties in reading. In D. Shankweiler & I. Y. Liberman (Eds.), *Phonology and reading disability: Solving the reading puzzle* (pp. 35–68). Ann Arbor: University of Michigan Press.
- Shu, H., Chen, X., Anderson, R. C., Wu, N., & Xuan, Y. (2003). Properties of school Chinese: Implications for learning to read. *Child Development*, 74, 27–47.
- Singer, M. (1994). Discourse inference processes. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 479–515). New York: Academic Press.
- Siok, W. T., & Fletcher, P. (2001). The role of phonological awareness and visual–orthographic skills in Chinese reading acquisition. *Developmental Psychology*, 37, 886–899.
- Spinks, J. A., Liu, Y., Perfetti, C. A., & Tan, L. H. (2000). Reading Chinese characters for meaning: The role of phonological information. *Cognition*, 76, B1–B11.
- Stothard, S. E., & Hulme, C. (1992). Reading comprehension difficulties in children: The role of language comprehension and working memory skills. *Reading and Writing: An Interdisciplinary Journal*, 4, 245–256.
- Swanson, H. L. (1992). Generality and modifiability of working memory among skilled and less skilled readers. *Journal of Educational Psychology*, 84, 473–488.
- Swanson, H. L., & Berninger, V. W. (1995). The role of working memory

- in skilled and less skilled readers' comprehension. *Intelligence*, 21, 83–108.
- Swanson, H. L., & Berninger, V. W. (1996). Individual differences in children's working memory and writing skill. *Journal of Experimental Child Psychology*, 63, 358–385.
- Swanson, H. L., Trainin, G., Necochea, D. M., & Hammill, D. D. (2003). Rapid naming, phonological awareness, and reading: A meta-analysis of the correlation evidence. *Review of Educational Research*, 73, 407–440.
- Tan, L. H., & Perfetti, C. A. (1998). Phonological code as early sources of constraint in Chinese word identification: A review of current discoveries and theoretical account. *Reading and Writing: An Interdisciplinary Journal*, 10, 155–164.
- Tan, L. H., & Perfetti, C. A. (1999). Phonological activation in visual identification of Chinese two-character words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 382–393.
- van den Bos, K. P., Zijlstra, B. J. H., & Spelberg, H. C. (2002). Life-span data on continuous-naming speeds of numbers, letters, colors, and pictured objects, and word-reading speed. *Scientific Studies of Reading*, 6, 25–49.
- van den Bos, K. P., Zijlstra, B. J. H., & van den Broeck, W. (2003). Specific relations between alphanumeric-naming speed and reading speeds of monosyllabic and multisyllabic words. *Applied Psycholinguistics*, 24, 407–430.
- van den Broek, P. (1994). Comprehension and memory of narrative texts: Inferences and coherence. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 539–588). New York: Academic Press.
- Wang, L. (1985). *Zhongguo xiandai yufa* [Modern Chinese grammar]. Beijing, China: Shangwu Yinshuguan.
- Williams, J. P., Lauer, K. D., Hall, K. M., Lord, K. M., Gugga, S. S., Bak, S.-J., et al. (2002). Teaching elementary school students to identify story themes. *Journal of Educational Psychology*, 94, 235–248.
- Zhang, S., & Perfetti, C. A. (1993). The tongue-twister effect in reading Chinese. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 1082–1093.

(Appendix follows)



## Appendix

## Sample Chinese Items With English Translation

## 1 閱讀理解 (Text comprehension No. 2 "Spear and Shield")

發明家手持矛和盾，與朋友比賽。

對方的矛如雨點般向他刺來，發明家用盾左抵右擋，還是難以招架。在這緊張危急的關頭，發明家忽然產生了一個想法：「盾太小了！如果盾大得像鐵屋子，我鑽在鐵屋子裏，敵人就一槍也戳不到我了！」

可是，這樣固然安全，自己卻變成了只能縮在殼裏保命的蝸牛或烏龜。

對了，在鐵屋子上開個小洞，從洞裏伸出「矛」——

槍口或砲口。當然，這鐵屋子還要會跑，得裝上輪子，安上履帶。於是，發明家發明了坦克。

The inventor was holding a spear and a shield and was jousting with his friend.

His opponent was using his spear to hit at him like raindrops. Even though the inventor was using his shield left and right to ward off the attack, he still found it difficult to defend himself. During this critical moment, the inventor had a thought: "The shield is too small! If it were like a big house made of iron, I can hide inside this iron house and the enemy cannot spear me!"

However, while the iron house might protect him from the spear attack he would be like a snail or a tortoise hiding in its shell just to be safe.

He had the bright idea of opening a small hole in the iron house and of showing only his "spear"—a gun or cannon. Of course, this iron house will need to be able to move fast on wheels with traction tires. Thus was born the idea of the prototype modern attack tanks.

## 1a 矛和盾分別有什麼作用？

What are the different uses [functions] of the spear and the shield?

## 1b 請說明坦克成為強大的武器的原因。

Please explain why the tank is a powerful weapon.

## 1c 請給這篇短文加上合適的標題，並解釋你的理由。

Please give the short passage a title, and explain your reasons.

## 2 假詞 (Pseudoword)

炮喻[pau3 jy6]

橫蘊[wan4 wan3]

潭砌[tam4 tsai3]

■遮[luŋ1 dze1]

僂瀑[wui1 buk6/bou6]

塌減[tap3 gaam2]

盞錦[dzan2 gam2]

遂促[sœy6 gwat9]

鋒棕[fun1 dzun1]

疼撤[tan4 tsit8]

## 3 「語文工作記憶」項目 Memory Span Task (MEMSP)

園子四面有高圍牆。 The courtyard has on four sides a high wall.

用顯微鏡才能看到細胞。 A microscope can show minute cells.

海底蘊藏著豐富的石油。 In the seabed can be found rich reserves of oil.

問題：顯微鏡有甚麼作用？ Question: What does a microscope do?

答案：放大（物體的影像） Answer: Amplify or magnify (images of objects).

詞語：圍牆、細胞、石油 Last word of each sentence: wall, cells, oil

## 4 急口令 (tongue-twister)

1. 東門南門 [dung1 mun4 naam4 mun4] East gate south gate

東門東家 [dung1 mun4 dung1 gaa1], East gate east (also host house),

南門董家 [naam4 mun4 dung2 gaa1]. South gate Tung house.

東董兩家 [dung1 dung2 loeng5 gaa1], East Tung [play on speech sound with mainly tones 1 and 2] two house(s).

同種冬瓜 [tung4 zung3 dung1 gwaal]. Together planting melon(s).

Appendix (*continued*)

## 5 Phonological subsyllabic segmentation

## Deletion of rime (DR)

斷[duan]	生[sheng]	棉[mian]
鵝[goose]	盒[box]	襯衫[shirt]

## Deletion of onset (DO)

消[xiao]	棕[zong]	蝦[xia]
五[five]	金[gold]	傳真[fax]

Received July 25, 2006

Revision received May 15, 2007

Accepted June 4, 2007 ■

**E-Mail Notification of Your Latest Issue Online!**

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at <http://notify.apa.org/> and you will be notified by e-mail when issues of interest to you become available!



# Development of Word Reading Fluency and Spelling in a Consistent Orthography: An 8-Year Follow-Up

Karin Landerl

University of Tuebingen and University of Salzburg

Heinz Wimmer

University of Salzburg

In a longitudinal study, development of word reading fluency and spelling were followed for almost 8 years. In a group of 115 students (65 girls, 50 boys) acquiring the phonologically transparent German orthography, prediction measures (letter knowledge, phonological short-term memory, phonological awareness, rapid automatized naming, and nonverbal IQ) were assessed at the beginning of Grade 1; reading fluency and spelling were tested at the end of Grade 1 as well as in Grades 4 and 8. Reading accuracy was close to ceiling in all reading assessments, such that reading fluency was not heavily influenced by differences in reading accuracy. High stability was observed for word reading fluency development. Of the dysfluent readers in Grade 1, 70% were still poor readers in Grade 8. For spelling, children who at the end of Grade 1 still had problems translating spoken words into phonologically plausible letter sequences developed problems with orthographic spelling later on. The strongest specific predictors were rapid automatized naming for reading fluency and phonological awareness for spelling. Word recognition speed was a relevant and highly stable indicator of reading skills and the only indicator that discriminated reading skill levels in consistent orthographies. Its long-term development was more strongly influenced by early naming speed than by phonological awareness.

**Keywords:** reading fluency, spelling, longitudinal assessment, long-term prediction of reading and spelling

Only recently, word reading fluency has moved from being a neglected aspect of reading to being a popular topic in the field of reading research. As Ehri (2002) stated,

One of the great mysteries to challenge researchers is how people learn to read and comprehend text rapidly and with ease. . . . A large part of the explanation lies in how they learn to read individual words. Skilled readers are able to look at thousands of words and immediately recognize their meanings without any effort. (p. 7)

A longitudinal study with a sample of normally developing readers of German allowed us to follow the development of word reading fluency and also orthographic spelling skills over almost 8 years, from school entry through to Grade 8.

Most longitudinal studies with English-speaking children did not even include measures of reading fluency but reported devel-

opment in word reading accuracy only (see Parrila, Aunola, Leskinen, Nurmi, & Kirby, 2005, for a review). In phonologically more transparent orthographies, however, word reading accuracy is often close to or at ceiling after only a few months of formal reading instruction (Cossu, Giuliotta, & Marshall, 1995; Oney & Goldman, 1984; Wimmer & Hummer, 1990), and word reading speed is the only word reading measure that differentiates between good and poor readers in higher grades.

There exist a few studies reporting longitudinal development of reading fluency in orthographies with consistent grapheme–phoneme correspondences. Bast and Reitsma (1998) followed the reading development of 235 Dutch children from kindergarten to the end of Grade 3. The first assessment of word reading fluency was carried out 3 months after the onset of formal reading instruction, with two further assessments during first grade, two assessments in second grade, and a final assessment toward the end of third grade. At each of the measurement points, children were given word lists, and the dependent variable was the number of words read correctly within a certain time limit (between 1 and 3 min). Unfortunately, Bast and Reitsma did not report reading accuracy separately from reading speed. Only for the later measurement points is it likely that accuracy among these readers of the consistent orthography of Dutch was high so that the reported measure can be interpreted as a pure measure of reading fluency. The finding of importance is that, with respect to reading fluency, the rank ordering of individuals was highly stable for the measurements after Grade 1, and individual differences in word recognition fluency even increased over time.

Parrila et al. (2005) recently reported a longitudinal assessment of a sample of 197 Finnish children's development of oral reading fluency (reading a short text as quickly as possible) from the

---

*Editor's Note.* Linda Baker served as the action editor for this article.—KRH

---

Karin Landerl, Department of Psychology, University of Tuebingen, Germany, and Department of Psychology, University of Salzburg, Austria; Heinz Wimmer, Department of Psychology, University of Salzburg.

The assessments in Grades 1 and 4 were funded by Austrian Science Fund Grant P12481-SOZ. We thank Christine Huber, Andreas Kaiser, Thomas Lackner, Elisabeth Raffl, and Agnes Schinwald for their help in data collection.

Correspondence concerning this article should be addressed to Karin Landerl, Department of Psychology, University of Tuebingen, Gartenstraße 29, 72074, Tuebingen, Germany. E-mail: karin.landerl@uni-tuebingen.de

beginning of Grade 1 until the end of Grade 2. This study also indicated high stability of reading fluency over four measurement points (two per school year), but, in contrast to Bast and Reitsma (1998), the authors did not find evidence that individual differences increased during the project period. Once again, it is problematic that Parrila et al. (2005) did not report reading accuracy separately from reading fluency, so that it is unclear whether these two measures were confounded. Parrila et al. (2005) also reported high stability in reading development for a sample of 198 Canadian children who were followed from Grade 1 up to Grade 5. However, this finding is not informative with respect to reading fluency development as—quite typical for studies with English-speaking children—the dependent measures only included measures of word reading accuracy but no measure of reading speed.

De Jong and van der Leij (2002) followed the reading fluency development of 141 Dutch children. They reported a correlation of .69 between word reading fluency (number of words read correctly in 1 min) at the end of Grade 1 and Grade 3, suggesting a high stability over the years. When the students were in Grade 6, de Jong and van der Leij (2003) retested specifically selected subgroups from their longitudinal sample. A group of dyslexic children was selected on the basis of their reading skills in Grade 3, when their level of word reading fluency was, on average, more than two standard deviations below the age norms. This strong deviation from the mean was still evident in Grade 6, confirming the high stability of word reading fluency. It is interesting that at the end of Grade 1, dyslexic children's word reading accuracy was high and not significantly different from that of a control group of children with age-adequate reading development. This finding indicates that the extremely low reading fluency that these children exhibited was not caused by lack of knowledge of the graphophonic connections.

In an Austrian longitudinal study Klicpera and Schabmann (1993) documented 356 German-speaking children's reading development from Grade 2 to Grade 8. In the middle of Grade 2, children were assigned to five reading level groups on the basis of their reading fluency. The stability for the highest and lowest reading level groups in particular was very high. Only 2 children who were among the 5% of poorest readers in Grade 2 developed average reading skills in Grade 8, but 94% of this group still performed below Percentile 15. The children who had been assigned to the lowest reading level in Grade 2 exhibited a reading fluency in Grade 8 that was comparable to that of average readers in Grade 2. Thus, their reading speed development was about 6 years delayed. Nevertheless, word reading accuracy was high even among the poorest reading level group. Already in Grade 2 they read fewer than 20% of the presented words incorrectly. This is impressive, given the fact that reading was assessed under speed instruction. Without the emphasis on reading as fast as possible, accuracy might have been even higher. Obviously, there was not much room for further improvement in reading accuracy later on. In Grade 8, even the poor readers made errors on fewer than 10% of the presented words.

The current study followed children's literacy development over an even longer time period than any of the studies reported above, namely from school entry until the middle of Grade 8. This longitudinal assessment allowed us to test the stability of reading fluency development and also of orthographic spelling skills over three assessments (end of Grade 1, beginning of Grade 4, and

middle of Grade 8). Furthermore, the design of the study also enabled us to assess the predictive power of standard psycholinguistic measures for the longitudinal development of reading fluency and spelling. The first assessment was carried out at the beginning of Grade 1, before the onset of formal reading instruction, and included measures of phonological processing (phonological awareness, phonological short-term memory), letter knowledge, and rapid automatized naming. It is important to note that the German/Austrian kindergarten system does not provide explicit reading preparation. During their preschool years, children are supposed to develop in the context of playing. Reading, spelling, and arithmetic are only introduced in first grade. This difference between the German and the Anglo-American preschool education systems became clearly evident in a direct comparison of the letter knowledge and phonological awareness of Austrian and U.S. kindergartners carried out by Mann and Wimmer (2002). Percentages of correctly named letters were above 90% for U.S. kindergartners but only around 30% for their Austrian age mates.

Findings of earlier phases of this longitudinal assessment were reported in Wimmer, Mayringer, and Landerl (2000, Study 2) and Wimmer and Mayringer (2002, Study 2). The main findings were that serious deficits in reading fluency and orthographic spelling were not strongly associated with each other and that they might have been based on different cognitive deficits. In Wimmer et al. (2000), we reported that those children who entered school with serious deficits in phonological awareness later on developed problems in orthographic spelling and irregular word reading but not in the domain of phonological decoding (i.e., nonword reading). Obviously, this is in striking contrast to findings with English-speaking children, for whom deficits in phonological awareness have been demonstrated to be closely associated with problems in phonological decoding (Vellutino, Fletcher, Snowling, & Scanlon, 2004) and for whom phonological decoding, in turn, is seen as a bottleneck to further reading acquisition (Perfetti, 1985). The combination of the highly consistent grapheme-phoneme correspondences of German orthography and a systematic phonics teaching approach seems to enable even children who enter school with low phonological skills to master the initial hurdle of phonological decoding. Only when the relationships between spoken and written language become more complex, as in irregular word reading or spelling, where German orthography is clearly less consistent than in the reading direction, do early phonological deficits seem to have a more persistent negative influence on children's literacy development.

While early phonological awareness deficits did not have a strong influence on children's later reading development, deficits in sequential naming speed, measured by a rapid automatized naming paradigm, did: Children who entered school with a serious naming speed deficit turned out to read both words and nonwords with strikingly low reading fluency in Grade 4. In addition, they showed problems in orthographic spelling.

In a second, retrospective analysis of the longitudinal data, Wimmer and Mayringer (2002) selected three deficit groups from the large normative sample of 301 children on the basis of their reading and spelling performance in Grade 4: A group of dysfluent readers who had age-adequate skills in orthographic spelling, a group of poor spellers with age-adequate reading fluency, and a group of students who were deficient in both reading fluency and orthographic spelling. It turned out that the dysfluent readers



already showed marked problems in rapid automatized naming of objects at school entry. The poor spellers, conversely, typically started into their school career with a deficit in phonological awareness. The group with combined reading and spelling deficits showed problems in both cognitive domains when they entered school.

In summary, the findings suggest that, at least in the German context of a consistent orthography and a synthetic phonics teaching approach, reading development depends more strongly on adequate naming speed than on phonological awareness. Similar findings come from studies in Dutch (de Jong & van der Leij, 1999, 2002) and Italian (Di Filippo et al., 2005). This is plausible, as the main problem of poor readers in phonologically transparent orthographies is extremely low reading fluency, while reading accuracy is usually high (de Jong & van der Leij, 1999; Landerl, Wimmer, & Frith, 1997; Wimmer, 1993; Zoccolotti et al., 2005). The literacy skill that does depend on good phonological awareness is spelling rather than reading. However, note that it is not the component of phoneme segmentation in spelling that is predicted by phonological awareness. After only few months of formal instruction, German-speaking children are usually very well able to translate even complex phoneme sequences into phonologically plausible grapheme sequences (Wimmer & Landerl, 1997). However, such a systematic phoneme–grapheme translation is not sufficient to spell words orthographically correctly, because the consistency of phoneme–grapheme correspondences is much lower than the consistency of grapheme–phoneme correspondences. Quite often, there are two or three different possibilities to translate a certain phoneme into a phonologically adequate grapheme, but obviously only one of these possibilities is orthographically correct. A plausible explanation for why phonological awareness might be relevant for orthographic spelling is that the build-up of stable representations in orthographic memory requires phonological underpinning—that is, multiple and redundant associations between the letters and sounds of specific words (Ehri, 1992; Perfetti, 1992). Deficits in phonological awareness might be an indicator that phonological representations are not well enough specified to be accessible for the formation of such associations (Snowling, 2000).

Two main questions are addressed on the basis of the Grade 8 assessment: First, how stable was the development of reading fluency and orthographic spelling over the 8 years of the reported longitudinal study? The second question we want to address is to what extent the pattern of predictions reported for Grade 4 can be replicated in Grade 8. Predictor measures assessed at the beginning of Grade 1 were tests of letter knowledge, rapid automatized naming, phonological processing (awareness and short-term memory), and nonverbal IQ.

After Grade 4, all children change to different secondary schools in the Austrian school system, which made it impossible to keep track of all participants of the originally large sample of 356 children. Still, we did not want to miss the unique opportunity to get longitudinal information on children's literacy development by collecting data from as many children as possible before some of them left school after the 9 years of obligatory schooling in the Austrian school system.

## Method

### *Sample and Procedures*

In Grade 1, a large sample of 356 children, from eight schools and 23 classrooms, was tested. In Grade 4, 296 children of the original sample could be retested; 60 children had moved or were not in school on the day of testing. Almost 8 years after the first assessment, we tried to contact all students who had participated in Grade 4. For 83 students, no addresses were available, so that they could not be invited to participate. From those who received our letter of invitation to participate in a further assessment at the Department of Psychology, 98 did not respond or declined. The number of students who agreed to participate was 115 (65 girls and 50 boys). All students attended Grade 8. Mean ages at the various assessment points were 6.9 years at the beginning of Grade 1, 7.5 years at the end of Grade 1, 10.2 years in Grade 4, and 14.3 years in Grade 8 ( $SD = 0.5$  years at all assessment points). Only data from children who participated in all assessments are reported in this article.

In the first 2 months after children entered school, we assessed phonological processing (awareness and short-term memory) and rapid automatized naming. Nonverbal IQ was measured in the second test session at the end of Grade 1. Reading and spelling skills were assessed at the end of Grade 1, beginning of Grade 4, and middle of Grade 8. All first-grade and fourth-grade assessments were carried out in the school; only the last assessment in Grade 8 took place in our lab. Apart from the spelling tests in Grade 4 and Grade 8, which were administered groupwise, all testing was done individually. Individual testing was carried out in a separate room and lasted between 30 and 40 min at each measurement point. All testing was carried out by trained master's and doctoral students.

### *Measures*

#### *Beginning of Grade 1*

**Letter knowledge.** Children were presented with the most frequent 19 uppercase letters and were asked to name them. Either letter name or letter sound counted as a correct response.

**Rapid automatized naming.** Four object pictures (car, ball, dog, mouse) were introduced, and the child was asked to name them as rapidly as possible. A further practice trial with two lines of these pictures familiarized children with the demands of rapid naming. The test page consisted of eight lines, with 4 pictures on each line (total of 32 pictures), with the pictures in varying order in each line. After the first four lines, the experimenter recorded the response time without interrupting the child. At the end of the eight lines, response time was recorded again. The correlation between the first and second halves of the naming speed task was .61.

**Phonological short-term memory.** Children were asked to repeat 16 sets of (C)CVC nonsense syllables. Eight items consisted of two syllables (e.g., *tes-bof*), the next eight syllables consisted of three syllables (e.g., *gat-fos-hap*). The correlation between two- and three-syllabic items was  $r(295) = .48, p < .001$ .

**Phonological awareness.** Manipulations on the phoneme level are usually rather difficult for Austrian 6-year-olds, who receive no reading preparation in kindergarten. In earlier studies, we found

floor effects for traditional phoneme awareness tasks such as phoneme deletion or phoneme substitution (Mann & Wimmer, 2002; Wimmer, Landerl, Linortner, & Hummer, 1991). To design a task with an acceptable level of difficulty, we developed a paradigm in which children had to imitate a phoneme segmentation modeled by the experimenter. The experimenter explained that words can be divided into smaller parts and demonstrated this with *fee*: /f/-/e:/. Children were instructed to imitate the segmentation procedure modeled by the experimenter. After the child had repeated the practice item correctly, the experimenter presented six CV words and six CVC words. For each item, the experimenter said first the word and then the constituent phonemes, and children had to imitate her by producing both the word and the segmented sounds. The idea was that a child who did not understand the relation between the word and its segments would find the task difficult, while awareness of the sound structure of words would make it easier to imitate the experimenter correctly. This measure bears similarity with the phonological memory task, in which nonsense syllables instead of phoneme sounds had to be repeated. However, the correlation between the phoneme task and the nonsense syllable task was only moderate (.30,  $p = .001$ ), so that only 9% of the variance of our phonological awareness measure could be explained by phonological memory capacity. The clearly higher correlation with letter knowledge (.57,  $p < .001$ ) suggests that children profited from knowledge of letter-sound associations. The correlation between letter knowledge and phonological memory was also only moderate (.20,  $p < .05$ ), indicating that children's letter knowledge was not only based on rote memory of the letter names. The reliability of the task was satisfactory. The correlation between CV and CVC items was .58 ( $p < .001$ ). For a subsample of 40 randomly selected children, we carried out an item-based analysis. Internal consistency turned out to be high ( $\alpha = .80$ ).

### End of Grade 1

**Nonverbal intelligence.** To assess children's nonverbal cognitive skills, we gave them Raven's Coloured Progressive Matrices (Schmidtke, Schaller, & Becker, 1978).

**Reading fluency.** Children were given two word and two nonword lists. One word list consisted of 9 one- and two-syllabic high-frequency concrete nouns; the other one consisted of 9 one-syllabic high-frequency function words. The task was introduced by a practice sheet with six words. One of the nonword lists consisted of 9 two-syllabic nonwords without consonant clusters; the other list included 9 one-syllabic nonwords with consonant clusters in onset and/or coda position. The nonword reading paradigm was also introduced by a practice list consisting of six nonwords. Children were instructed to read "as quickly as possible, without making mistakes." Both number of reading errors and reading time for each list were recorded: Reliability of the fluency tasks (syllables per minute) is evident from the high correlations among the four measures (ranging from .83 to .92,  $ps < .001$ ).

**Spelling.** Eleven high-frequency words were dictated to each child individually. After only 8 months of formal instruction, it did not seem appropriate to expect children to spell the words orthographically correctly. We rather wanted to know how well children would be able to translate the phoneme code into a phonologically plausible letter sequence. Thus, we scored whether a spelling was phonologically adequate in the sense that each sound in the word pronunciation was translated by an acceptable grapheme. To make

phoneme segmentation more difficult, we ensured that each of the words included a consonant cluster in the word onset position, and three words included a second consonant cluster in the word-final position.

### Grade 4

**Reading fluency.** Children were asked to read aloud a short story and two word lists from a standardized reading test battery (Landerl, Wimmer, & Moser, 1997). The text consisted of 57 words and was of simple content, so that reading time would not be affected by comprehension difficulties. The word lists each consisted of 11 complex compound words typical for German language (e.g., *Fruchtsaft*—fruit juice, *Geburtstagskuchen*—birthday cake). Children were instructed to read "as quickly as possible, without making mistakes." The correlation between reading fluency for the short text and the compound word lists was .88 ( $p < .001$ ). The test handbook reports satisfactory reliability (parallel test method) for both compound words (.93) and text (.90).

**Spelling.** Children had to fill 35 dictated words into sentence frames clarifying the word meaning. The words were specifically selected so that a simple phoneme-grapheme translation would not be sufficient to spell the words orthographically correctly. Each word included at least one orthographic marker typical for German orthography. The full sentence was read out by the experimenter, then the word to be spelled was repeated. Finally, the full sentence was repeated once more. Half of the sentences were taken from a standardized spelling test (Landerl, Wimmer, & Moser, 1997) with adequate reliability (.74); no reliability data were available for the current sample.

### Grade 8

**Reading fluency.** Reading fluency was assessed with two text reading paradigms. All texts were short and simple, so that good comprehension was ensured. The main criterion was the speed with which the texts could be read. First, participants were asked to read two short texts out loud. Text 1 was about the development of the Internet and consisted of 73 words and 177 syllables. Text 2 was about a new electric fence for cattle and consisted of 92 words and 156 syllables. Both errors and reading times were recorded. The correlation for reading time for the two texts was .85 ( $p < .001$ ). Asking participants to read aloud is an efficient way to assess both reading accuracy and reading speed; however, reading aloud is a rather untypical reading activity for older readers. The more natural reading activity is silent reading, which was also assessed. Three different texts, consisting of 260 words and 536 syllables altogether, were presented on a computer screen. To cover the probably widespread interests of our participants, we ensured that the texts were about very different topics (the Chinese Wall, legal restrictions for selling certain alcoholic drinks to adolescents, and an incident in which a hawk attacked a dog). Two texts were selected from daily newspapers, and one text came from a science book for adolescents. Participants were asked to read each text silently as quickly as possible and to press the space bar, triggering the response time measurement, when finished. The procedure was introduced by a short practice text consisting of 57 words. To ensure that participants read the texts carefully, the experimenter asked a short question after each text. Although they



had not been informed about the comprehension questions before reading the text, all participants answered these questions correctly. These correct answers show that participants attended to the text and that the texts were of a rather simple comprehension level for that age group. It is therefore unlikely that reading fluency was strongly influenced by comprehension difficulties. Obviously, word reading accuracy could not be measured for this paradigm; only reading time was recorded. Correlations between the texts ranged from .88 to .91 ( $ps < .001$ ).

*Spelling.* A standardized classroom test (Kersting & Althoff, 2004) was given in which students had to supply 68 dictated words or phrases to a cloze text. Each sentence was read out loud by the experimenter, and the word to be spelled was repeated. The test handbook reports high internal consistency (Cronbach's  $\alpha = .93$ ).

Results

Attrition

Table 1 presents the means and standard deviations for the prediction measures separately for those 115 children who still participated in the study in Grade 8 and those 241 children who declined to participate in Grade 8. For letter knowledge, rapid automatized naming, phonological short-term memory, and nonverbal IQ, no statistical differences were found. On the phonological awareness measure, those children who dropped out of the study later on performed significantly lower than those who still participated in Grade 8,  $t(352) = 2.68, p = .008$ . Note, however, that the difference between the two groups was only 9%, which is small in relation to the standard deviation of 29%. We also compared Grade 4 performance for those 115 children who participated in all assessments and those 180 children who still participated in Grade 4 but not later on: No difference was found for word reading fluency (Grade 8 participants:  $M = 174$  syllables/min,  $SD = 47$ ; Grade 8 nonparticipants:  $M = 164$  syllables/min,  $SD = 47$ ),  $t(293) = 1.76, ns$ . Only with respect to spelling was there a rather small but, because of the large sample size, nevertheless significant difference: Children who dropped out of the study later on spelled 71% of the dictated words correctly, while those who participated in Grade 8 produced 75% correct spellings,  $t(293) = 2.1, p = .04$ .

Table 1  
*Descriptive Statistics of Prediction Measures for Children Who Participated in Grade 8 and Those Who Dropped Out of the Study*

Prediction measure	Participants ( <i>n</i> = 115)		Nonparticipants ( <i>n</i> = 241)	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Letter knowledge (% correct)	60.4	30.7	54.6	27.9
Phonological awareness (% correct)	60.5	28.5	51.6	29.4
Phonological short-term memory (% correct)	56.7	29.7	55.7	30.0
RAN (words/min)	56.1	12.1	54.8	11.1
Nonverbal IQ	117.0	12.5	115.5	12.4

Note. RAN = rapid automatized naming.

Development of Reading Accuracy

The upper section of Table 2 presents the accuracy measures for reading for the end of Grade 1, Grade 4, and Grade 8. It is evident that after only 1 year of formal reading instruction, children showed high accuracy for both words and nonwords. Actually, 51% and 33% of the children read the word and nonword lists without errors. Only 4 children (4% of the sample) read fewer than 80% of the words correctly, and only 12 children (10%) read fewer than 80% of the nonwords correctly. The minimum accuracy scores of 72% and 61% correct indicate that none of the young readers showed serious problems with phonological decoding. Although individual children certainly still needed more practice, all children showed good competence in translating grapheme sequences into word or nonword pronunciations; thus, they had good knowledge of the alphabetic mapping system. For the interpretation of all accuracy scores presented in Table 2, it must be noted that all reading tasks were given under time pressure. Children were instructed to “read as fast as you can, without making mistakes.” It can be assumed that accuracy scores would be even higher if the instruction had been to read as carefully as possible.

Development of Reading Fluency

For each grade level, we computed scores of number of syllables read per minute. Because reading accuracy was high in all assessments, it was not necessary to correct the reading fluency scores for number of errors, and the scores were not confounded with reading accuracy. In Grade 1, separate scores were computed for the two word reading lists and the two nonword reading lists. In Grade 4, the short text and the two word lists were combined for this measure. In Grade 8, separate scores were computed for reading aloud and reading silently. The descriptive statistics for these combined fluency measures are presented in the middle section of Table 2. It is not surprising that reading fluency increased dramatically over the 8 years of the study. In Grade 1, there was a very high correlation ( $r = .91, p < .001$ ) between reading fluency for words and nonwords, indicating that children were using a strategy of phonological decoding for words as well as nonwords. However, Table 2 shows that, even at that early point in reading development, children were clearly better at reading words compared to nonwords,  $t(114) = 14.9, p < .001$ . Another important observation is the enormous variability of the fluency measures that was evident in all grade levels. In Grade 8, silent reading was clearly faster than reading aloud. The correlation between reading aloud and silent reading was substantial ( $r = .71, p < .001$ ).

Development of Spelling Skills

The important finding here is that after only 1 year of formal instruction, children were highly accurate in translating word pronunciations into grapheme sequences. The lower section of Table 2 shows that in Grade 4 children spelled a larger percentage of the dictated words correctly than in Grade 8. This was due to the higher degree of difficulty of the word material used in the Grade 8 test. Kolmogorov–Smirnov tests showed that scores on both the Grade 4 and the Grade 8 spelling measures were normally distributed ( $Zs = 1.1$  and  $0.9$ , respectively,  $ps > .1$ ). The Grade 8 measure was a standardized spelling test, which allowed us to

Table 2  
Descriptive Statistics of Reading and Spelling Measures in  
Grades 1, 4, and 8

Measure	M	SD	Min	Max
Reading accuracy (% correct)				
End of Grade 1				
Words	95.4	6.2	72.0	100.0
Nonwords	92.2	8.8	61.0	100.0
Grade 4				
Compound words	96.9	4.9	73.0	100.0
Text	97.9	3.9	82.0	100.0
Grade 8				
Reading aloud	96.7	3.6	68.0	100.0
Word reading fluency (syllables per minute)				
End of Grade 1				
Words	71.9	34.6	18.6	175.6
Nonwords	46.8	20.5	11.2	137.3
Grade 4	174.5	47.5	59.0	285.4
Grade 8				
Reading aloud	267.2	50.7	45.0	396.4
Silent reading	486.8	158.3	64.7	984.9
Spelling (% correct)				
End of Grade 1				
(phonologically adequate)	93.4	10.9	36.4	100.0
Grade 4	75.3	16.1	22.9	100.0
Grade 8	53.7	19.0	0.0	97.1

compare our sample with the norming sample. The mean performance of the current sample corresponded to the 54th percentile.

### Stability of Reading Fluency and Spelling From Grade 1 to Grade 8

The second and third sections of Table 3 inform on the stability of the differences in reading fluency and spelling. In spite of the long time periods between the assessments, high correlations could be

observed for word reading fluency. Even the correlations between reading fluency in Grade 1 and Grade 8, with an interval of almost 7 years, were around .60. The correlations for reading aloud and reading silently in Grade 8 were almost identical. The correlation between phonological spelling in Grade 1 and orthographic spelling in Grade 1 was also substantial and close to .50. The correlation between orthographic spelling in Grade 4 and Grade 8 was once again high.

Figure 1 examines the stability of reading fluency in more detail by plotting word reading fluency in Grade 1 against word reading fluency (for reading aloud) in Grade 8. Almost all of the 11 children who performed more than one standard deviation below the group mean in Grade 1 were still more than one standard deviation below the mean in Grade 8, which indicates high stability for deficits in reading fluency. Only 1 of the group of novice poor readers showed a Grade 8 reading fluency score that was actually slightly above average; 2 further students moved to the low average range. However, children who were already reading fluently in Grade 1 (more than one standard deviation above the group mean) without exception developed at least average reading skills later on. Thus, they had a very low risk of developing reading problems over the following years. The same scatterplot was also inspected for silent reading in Grade 8, and the findings were highly similar. We also inspected the scatterplots for reading fluency in Grades 1 and 4 and in Grades 4 and 8, and the stability was similarly high.

In an analogous scatterplot, we inspected whether early phonological spelling—that is, systematic phoneme-grapheme translation—is a precursor of later orthographic spelling skills. In Figure 2, phonological spelling skills in Grade 1 are plotted against orthographic spelling skills in Grade 8. In the phonological spelling task in Grade 1, 68 children—that is, almost 70% of the sample—performed at ceiling (no spelling errors or one error). The interesting finding here is that most of the 15 children who had problems with phonological spelling at the end of Grade 1 and performed more than one standard deviation below the group mean developed below average orthographic spelling skills later on. Only 3 of these students showed orthographic spelling skills in the low average range in Grade 8.

Table 3  
Correlations Among Predictor Measures (Grade 1) and Reading Fluency and Spelling Measures (Grades 1, 4, and 8)

Measure	2	3	4	5	6	7	8	9	10	11	12
Grade 1 predictor measure											
1. Nonverbal IQ	.14	.15	.15	.09	.14	.23*	.32**	.30**	.31**	.30**	.43***
2. Letter knowledge	—	.57***	.20*	.39**	.40***	.43***	.32**	.35***	.25**	.47***	.47***
3. Phonemic awareness		—	.30**	.33***	.41***	.36***	.31**	.28**	.21*	.51***	.48***
4. Phonemic STM			—	.09	-.05	.02	.02	.00	.18	.23*	.28**
5. RAN				—	.45***	.46***	.34***	.34**	.31**	.35***	.32***
Reading fluency											
6. Grade 1					—	.69***	.59***	.64***	.35***	.52***	.52***
7. Grade 4						—	.81***	.77***	.40***	.63***	.69***
8. Grade 8: aloud							—	.71***	.48***	.59***	.64***
9. Grade 8: silent								—	.43***	.58***	.61***
Spelling											
10. Grade 1									—	.44***	.47***
11. Grade 4										—	.77***
12. Grade 8											—

Note. STM = short-term memory; RAN = rapid automatized naming.

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .



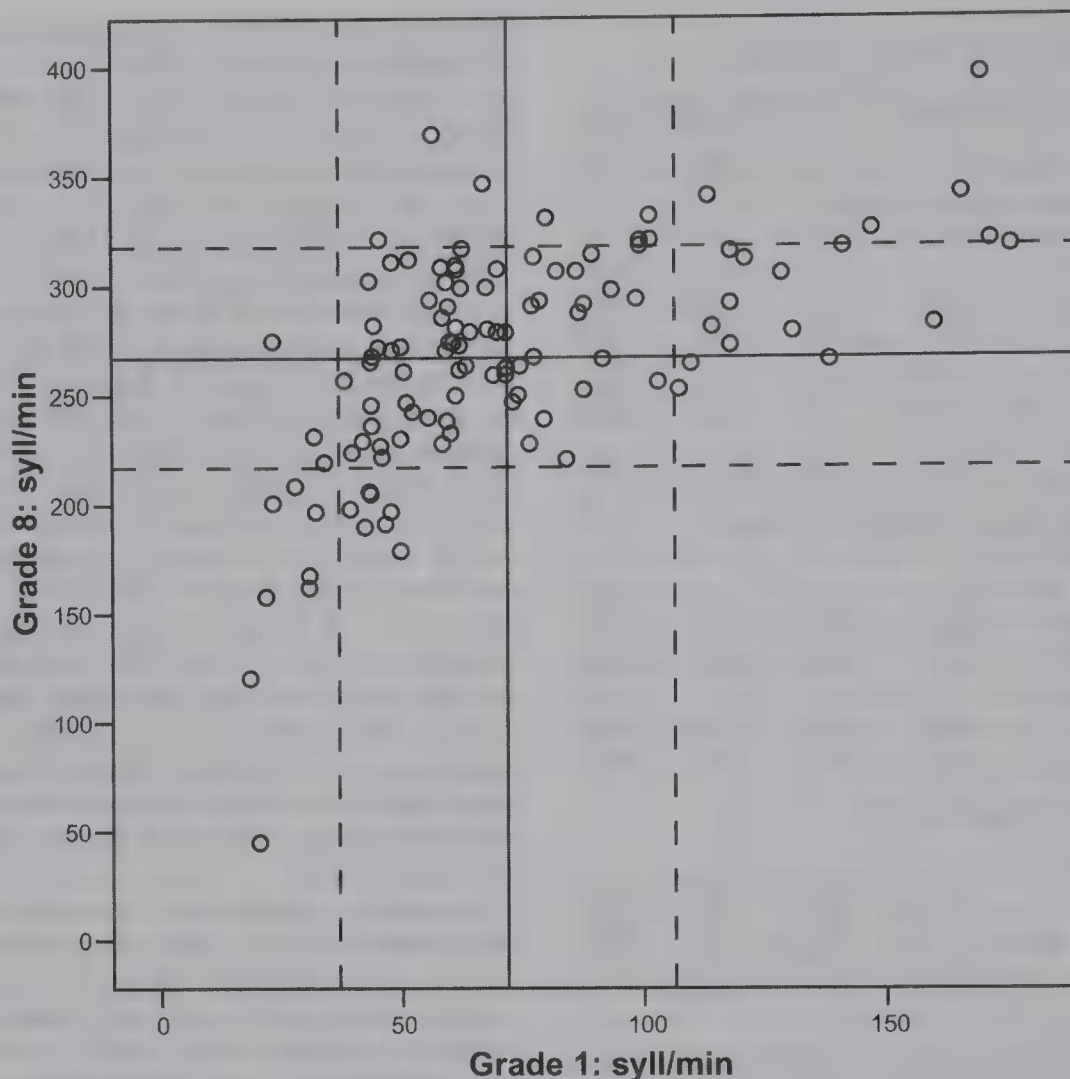


Figure 1. Scatterplot for reading fluency (syllables per minute) in Grades 1 and 8 (full reference lines represent the mean scores; dotted reference lines represent  $\pm 1$  SD).

### *Prediction of Reading Fluency and Orthographic Spelling*

The first section of Table 2 presents the correlations among the prediction measures and between these measures and the later reading fluency and spelling measures. Of interest is that nonverbal IQ was not associated with any of the other prediction measures but showed low but significant associations with reading fluency (not Grade 1) and also with all spelling measures. As already mentioned, early letter knowledge was substantially associated with phonological awareness and also with all other verbal predictor measures. Both letter knowledge and phonological awareness were associated with all reading fluency and spelling measures, and this was also the case for the rapid automatized naming measure. It is interesting that the phonological short-term memory measure was not associated with any of the reading fluency and spelling measures except for spelling in Grade 8.

To determine the specific contribution of rapid automatized naming and phonological processing (awareness and short-term memory) for the later literacy skills, we conducted two hierarchical linear regression analyses involving three steps each for each measure of reading fluency (Grades 1, 4, and 8) and orthographic spelling (Grades 4 and 8). Phonological spelling at the end of Grade 1 was not considered as a dependent measure because of the obvious ceiling effect in children's performance. In both analyses,

nonverbal IQ and letter knowledge entered the regression equation in the first step. Rapid automatized naming and phonological processing (awareness and short-term memory) entered the equation in either the second or the third step: In one of the analyses, rapid automatized naming entered the equation in the second step and phonological processing (awareness and short-term memory) entered in the third; the order of steps was reversed for the other analysis. Phonological awareness and phonological short-term memory were entered together, as our phonological awareness measure was obviously influenced by phonological memory. With these regression analyses, we hoped to determine whether each one of the predictor skills contributed variance to performance on the dependent measure that was not already accounted for by the other.

As can be seen from Tables 4 and 5, rapid automatized naming contributed independent variance to reading fluency in all grade levels. The phonological measures contributed significantly to word reading fluency in Grade 1 only, as well as to the spelling measures in Grades 4 and 8, but they were no longer predictive for the later assessments of reading fluency in Grades 4 and 8. In summary, these regression analyses indicate that the phonological measures (phonological awareness and phonological short-term memory) accounted for significant variance for early reading fluency (Grade 1) as well as for orthographic spelling, even if the

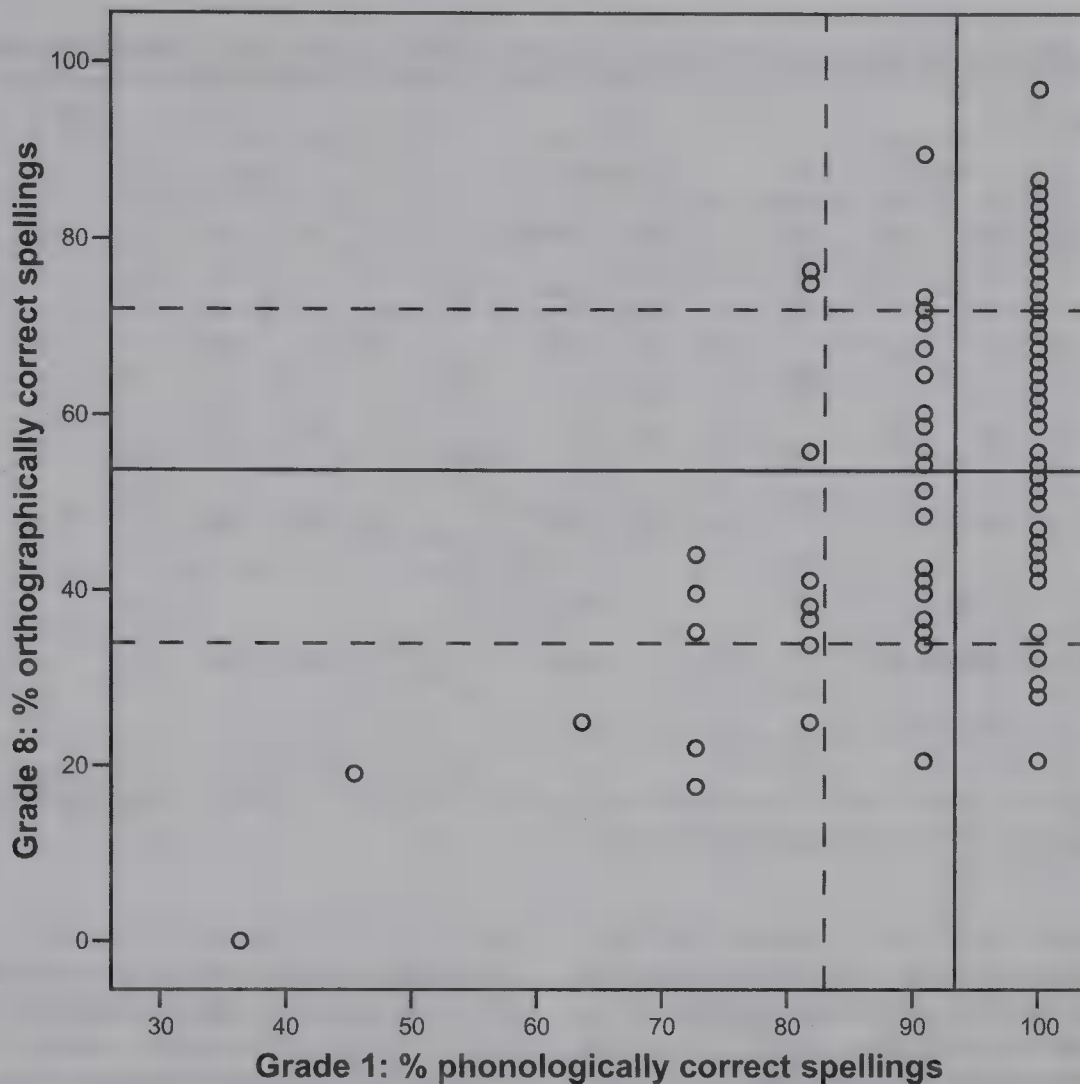


Figure 2. Scatterplot for phonological spelling in Grade 1 and orthographic spelling in Grade 8 (full reference lines represent the mean scores; dotted reference lines represent  $\pm 1$  SD).

variance contributed by rapid automatized naming was already accounted for. Early rapid automatized naming, conversely, was a significant and consistent predictor of reading fluency at all measurement points.

### Discussion

In line with earlier studies on the longitudinal development of reading fluency in consistent orthographies, we found a very high stability of reading fluency development over almost 8 years. This stability was evident from the strong correlations between the reading fluency measures in Grades 1, 4, and 8 and from the scatterplot presented in Figure 1. Fully 70% of the children who showed exceptionally slow and laborious reading (more than one standard deviation below the group mean) at the end of Grade 1 still were among the poorest readers of the current sample almost 8 years later. Only 3 children of the very poor first graders managed to develop at least (low) average reading skills later on. Not a single student of this group became an above average reader. Note that at the end of first grade, even the poor readers were well able to carry out the process of phonological decoding. Reading accuracy for words was high: Even the poorest readers were able to decode at least 60% of the nonwords correctly, and errors typically consisted of misreadings of only one letter. It is highly

likely that reading accuracy would have been even higher if the reading tasks had been presented without time pressure. The same word and nonword reading lists were previously used in a European cross-linguistic comparison of reading accuracy at the end of Grade 1 (Seymour, Aro, & Erskine, 2003). In this study, reading material matched for length, frequency, and syllable structure was given to children in 13 different orthographies. Apart from the Scottish and Danish groups (both phonologically rather opaque orthographies), all first-grade samples were at or close to ceiling for both words and nonwords, which confirms that reading accuracy development proceeds rapidly in orthographies with consistent grapheme–phoneme correspondences. The high accuracy in the current sample is actually a methodological advantage, as reading fluency development can be examined independently of accuracy problems.

Similar findings of high reading accuracy in association with a high stability of reading fluency were reported by Klicpera and Schabmann (1993) for another German-speaking sample and by de Jong and van der Leij (2002, 2003) for a Dutch sample. It is often assumed that deficits in word recognition are ultimately due to a delay in the acquisition of phonological decoding skills. Such a delay would prevent children from applying a so-called self-teaching mechanism (Share, 1995) by which unknown words are



Table 4  
Summary of Hierarchical Regression Analyses for Variables Predicting Word Reading Fluency in the Different Grade Levels

Variable	Grade 8															
	Grade 1				Grade 4				Aloud				Silent			
	R <sup>2</sup>	ΔR <sup>2</sup>	ΔF	β	R <sup>2</sup>	ΔR <sup>2</sup>	ΔF	β	R <sup>2</sup>	ΔR <sup>2</sup>	ΔF	β	R <sup>2</sup>	ΔR <sup>2</sup>	ΔF	β
Model 1																
Step 1	.15	.15	9.96***		.21	.21	15.05***		.18	.18	12.01***		.19	.19	12.79***	
Nonverbal IQ				.10				.17*				.27**				.26**
Letter knowledge				.14				.22*				.13				.20
Step 2	.23	.08	5.36**		.24	.03	2.17		.21	.03	2.17		.21	.02	1.65	
Phonemic awareness				.24*				.13				.16				.10
Phonemic STM				-.21*				-.12				-.12				-.13
Step 3	.30	.08	12.47**		.33	.09	14.41***		.25	.04	5.80*		.25	.04	5.67*	
RAN				.31**				.33***				.22*				.22*
Model 2																
Step 2	.25	.10	14.03***		.31	.10	15.98***		.22	.05	6.91*		.23	.04	6.39*	
RAN				.31**				.33***				.22*				.22*
Step 3	.31	.06	4.67*		.33	.02	1.56		.25	.02	1.67		.25	.02	1.34	
Phonemic awareness				.24*				.13				.16				.10
Phonemic STM				-.21*				-.12				-.12				-.13

Note. STM = short-term memory; RAN = rapid automatized naming.  
\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

decoded a few times until they can be stored in orthographic memory. For the current sample, however, no serious delays in the acquisition of phonological decoding were evident by the end of Grade 1. However, although more or less any grapheme sequence can be decoded with reasonable accuracy, some children could not make use of this self-teaching mechanism as efficiently as the rest of the sample. For these children, Ehri's (2002) statement that "reading the word a few times secures its connections in memory" (p. 11) is not appropriate.

In the current study, we also examined the longitudinal development of spelling skills and found a strong relationship between early problems in phonological spelling units and later orthographic spelling skills. This finding is relevant in two respects. First, it shows that even in the phonologically transparent orthography of German, good competence in translating phonemic segments into a grapheme sequence is indispensable. The main difference from findings in the phonologically much less transparent orthography of English is probably that successful word identification via phonological decoding acts as a learning mechanism, so that understanding how the orthography maps onto spoken words is much easier to acquire and does not pose a serious hurdle to the majority of children. Second, on first glance it is not obvious why early phonological skills should have an influence on orthographic spelling. Orthographic markers in German are only partly phonology based. The different spellings of homophone pairs such as *fiel* (*he fell*) and *viel* (*many*) or *Wahl* (*election*) and *Wal* (*whale*) are obviously not phonologically motivated, as the word pronunciations are identical. Still, children with poor phonological skills seem to be at a serious disadvantage to store orthographic spellings in memory. It is, of course, possible that phonological problems are only an indicator for more general verbal deficits, so that children who show phonological deficits at the beginning of liter-

acy instruction also have deficits in morphosyntactic skills that are relevant in spelling. Another explanation for the relation between early phonological skills and later spelling development is that the build-up of orthographic representations requires the formation of multiple associations between written and spoken words (Ehri, 1992; Perfetti, 1992). Children with phonological deficits might not be able to establish such multiple associations.

In addition to these findings on the long-term stability of reading fluency and spelling, the design of our longitudinal study also allowed us to examine the pattern of early prediction of phonological and naming speed measures for reading and spelling development over the full 8-year period of the study. So that readers interpret the findings adequately, it should once again be mentioned that the phonological awareness task applied in the present study was different from standard measures used in this field. Phonological awareness refers to the explicit access of the sublexical sound structure and is typically assessed by measures that require segmentation or manipulation of sound segments. However, earlier studies in our lab indicated that tasks requiring sound segmentation or manipulation are very difficult for German-speaking children before the onset of reading instruction (Mann & Wimmer, 2002), probably because sound games and reading preparation are largely absent from the preschool system in German-speaking countries. The imitation paradigm applied in the present study showed an adequate level of difficulty and therefore fulfilled the statistical requirements for the regression analyses presented here. An obvious concern is that rote memory of the critical word and its segments may boost performance. However, we note that the correlation between our phonological awareness measure and the phonological short-term memory measure was low. Furthermore, the short-term memory measure was not associated with any of the subsequent reading fluency measures or with the phonolog-

Table 5  
Summary of Hierarchical Regression Analyses for Variables Predicting Orthographic Spelling in Grades 4 and 8

Variable	Grade 4				Grade 8			
	$R^2$	$\Delta R^2$	$\Delta F$	$\beta$	$R^2$	$\Delta R^2$	$\Delta F$	$\beta$
Model 1								
Step 1	.28	.28	21.46***		.36	.36	31.51***	
Nonverbal IQ				.20**				.34***
Letter knowledge				.20*				.23*
Step 2	.35	.08	6.67**		.42	.06	5.50**	
Phonemic awareness				.30**				.22*
Phonemic STM				.05				.10
Step 3	.37	.02	3.17		.43	.01	2.30	
RAN				.15				.12
Model 2								
Step 2	.30	.03	4.63*		.38	.02	3.32	
RAN				.15				.12
Step 3	.37	.07	5.86**		.43	.05	4.94**	
Phonemic awareness				.30**				.22*
Phonemic STM				.05				.10

Note. STM = short-term memory; RAN = rapid automatized naming.  
\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

ical spelling measure in Grade 1. The measure was based on the idea that imitating the experimenter should be easier for children who understand that the experimenter is segmenting the simple words presented into functional sound units (i.e., phonemes).

In spite of the unconventional measure of phonological awareness, overall, earlier findings in German, Dutch, and Italian (de Jong & van der Leij, 1999, 2002; Di Filippo et al., 2005; Wimmer & Mayringer, 2002; Wimmer et al., 2000) can be confirmed. While the phonological measures (phonological awareness and phonological short-term memory) dropped out as significant predictors after first grade, naming speed assessed at school entry was a consistent predictor of reading fluency up to Grade 8. These findings seem to be different from findings with English-speaking children, for whom phonological awareness is often reported to be the best predictor of reading development (Cronin & Carver, 1998; Mann & Wimmer, 2002; Torgesen, Wagner, Rashotte, Burgess, & Hecht, 1997). It is possible that this finding was due to the fact that our imitation paradigm is not a phonological awareness task in the narrow sense—that is, children were not required to segment the phonemes of the presented words independently. Another explanation for this difference in findings might be that English studies usually do not assess the development of reading fluency but focus on the development of reading accuracy only. However, Kirby, Parrila, and Pfeiffer (2003) recently reported for a Canadian sample that phonological awareness assessed in kindergarten ceased to be a relevant predictor for word reading accuracy after Grade 2, while the prediction of kindergarten rapid automatized naming was only significant from Grade 3 on. It is thus possible that the same predictive patterns as reported here are also valid for English but that, as the acquisition of competent phonological decoding skills takes about 2 to 3 years longer in the phonologically opaque English orthography than in transparent orthographies such as German (Aro & Wimmer, 2003; Seymour et al., 2003), the change

in predictive strength of phonological awareness and rapid automatized naming takes place later in English than in German.

A methodological limitation of the present findings is the high attrition rate. Only about 35% of the children of the first grade sample still participated in the final assessment. One third of the sample was lost because of missing addresses, but another third was located and chose not to participate. The most likely selection bias for the final assessment is that adolescents with poor literacy skills might have tended to avoid the embarrassment of an assessment of their reading and spelling skills and therefore declined. If this is the case, this effect would certainly restrict the representativeness of our findings. Given the fact that we found no major differences between Grade 8 participants and nonparticipants on the prediction measures, our findings may overestimate the predictive power of our prediction measures if, indeed, the lower end of literacy skills was underrepresented in the Grade 8 sample. In other words, if it was the poorer readers who decided not to participate in Grade 8, they did not have lower scores on the prediction measures than the possibly better readers who did participate. Still, we do not think that such a selective attrition has been going on in our sample, as on the Grade 8 standardized spelling test the average performance corresponded to the 54th percentile. If the lower end of the distribution of literacy skills were missing from our sample, the mean level of performance could be expected to be above average.

In conclusion, the findings of the current longitudinal study, which followed German-speaking children's development of reading fluency and orthographic spelling from the beginning of Grade 1 until Grade 8, confirm once more that for children who start into literacy development with certain risk factors that can be identified easily, the long-term prognosis is strikingly poor. Sometimes children are able to compensate for inefficient word reading skills if they have sufficient time (Walczyk, Marsiglia, Johns, & Bryan,



2004), and therefore reading comprehension is not deleteriously affected. However, the correlation between reading fluency and reading comprehension was still substantial (.64) for the Austrian sample of 15-year-olds participating in the Program for International Student Assessment (Landerl & Reiter, 2002), which indicates that a high percentage of dysfluent readers also experience serious comprehension deficits.

A final comment seems justified on the issue of intervention. We did not assess whether or what kind of intervention the poorer readers in our sample received over the years. If they did receive intervention, it did not induce serious improvements of their reading fluency. Most reading intervention programs today focus on phonological awareness and phonological decoding in reading. Given that phonological awareness is not strongly related to the development of reading fluency and that even the poorer readers of the present study were well able to phonologically decode words and nonwords already at the end of Grade 1, such intervention programs do not seem adequate for the problems these children experienced. Indeed, current evaluations show that such phonology-based programs have positive effects for children with low reading accuracy but that they do not have an influence on these children's low reading fluency (Torgesen, Rashotte, & Alexander, 2001). Serious research efforts are needed to develop and evaluate intervention programs that are specifically tailored to the problems of these dysfluent readers with high reading accuracy. Attempts in that direction (e.g., Levy, 2001; Thaler, Ebner, Wimmer, & Landerl, 2004) show that dysfluent reading is not only a highly stable characteristic but also one that is hard to remediate.

## References

- Aro, M., & Wimmer, H. (2003). Learning to read: English in comparison to six more regular orthographies. *Applied Psycholinguistics*, 24, 621–635.
- Bast, J., & Reitsma, P. (1998). Analyzing the development of individual differences in terms of Matthew effects in reading: Results from a Dutch longitudinal study. *Developmental Psychology*, 34, 1373–1399.
- Cossu, G., Giuliotta, M., & Marshall, J. C. (1995). Acquisition of reading and written spelling in a transparent orthography: Two non-parallel processes? *Reading and Writing*, 7, 9–22.
- Cronin, V., & Carver, R. (1998). Phonological sensitivity, rapid naming, and beginning reading. *Applied Psycholinguistics*, 19, 447–461.
- de Jong, P. F., & van der Leij, A. (1999). Specific contributions of phonological abilities to early reading acquisition: Results from a Dutch latent variable longitudinal study. *Journal of Educational Psychology*, 91, 450–476.
- de Jong, P. F., & van der Leij, A. (2002). Effects of phonological abilities and linguistic comprehension on the development of reading. *Scientific Studies of Reading*, 6, 51–77.
- de Jong, P. F., & van der Leij, A. (2003). Developmental changes in the manifestation of a phonological deficit in dyslexic children learning to read a regular orthography. *Journal of Educational Psychology*, 95, 22–40.
- Di Filippo, G., Brizzolara, D., Chilosi, N., De Luca, M., Judica, A., Pecini, C., et al. (2005). Rapid naming, not cancellation speed or articulation rate, predicts reading in an orthographically regular language (Italian). *Child Neuropsychology*, 11, 349–361.
- Ehri, L. C. (1992). Reconceptualizing the development of sight word reading and its relationship to recoding. In P. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 107–143). Hillsdale, NJ: Erlbaum.
- Ehri, L. C. (2002). Phases of acquisition in learning to read words, and implications for teaching. *British Journal of Educational Psychology: Monograph Series*, 2(Serial No. 1), 7–28.
- Kersting, M., & Althoff, K. (2004). *Rechtschreibungstest (RT)* [Spelling test]. Göttingen, Germany: Hogrefe.
- Kirby, J. R., Parrila, R., & Pfeiffer, S. L. (2003). Naming speed and phonological awareness as predictors of reading development. *Journal of Educational Psychology*, 80, 437–447.
- Klicpera, C., & Schabmann, A. (1993). Do German-speaking children have a chance to overcome reading and spelling difficulties? A longitudinal survey from the second until the eighth grade. *European Journal of Psychology of Education*, 8, 307–323.
- Landerl, K., & Reiter, C. (2002). Lesegeschwindigkeit als Indikator für basale Lesefertigkeiten [Reading fluency as an indicator for basic reading skills]. In C. Wallner-Paschon & G. Haider (Eds.), *PISA PLUS 2000: Thematische Analysen nationaler Projekte* (pp. 61–66). Innsbruck, Austria: Studien Verlag.
- Landerl, K., Wimmer, H., & Frith, U. (1997). The impact of orthographic consistency on dyslexia: A German–English comparison. *Cognition*, 63, 315–334.
- Landerl, K., Wimmer, H., & Moser, E. (1997). *Salzburger Lese- und Rechtschreibtest* [Salzburg Reading and Spelling Test]. Bern, Switzerland: Hans Huber.
- Levy, B. A. (2001). Moving the bottom: Improving reading fluency. In M. Wolf (Ed.), *Dyslexia, fluency, and the brain* (pp. 357–379). Timonium, MD: York Press.
- Mann, V., & Wimmer, H. (2002). Phoneme awareness and pathways into literacy: A comparison of German and American children. *Reading and Writing*, 17, 653–682.
- Oney, B., & Goldman, S. R. (1984). Decoding and comprehension skills in Turkish and English: Effects of the regularity of grapheme–phoneme correspondences. *Journal of Educational Psychology*, 76, 557–568.
- Parrila, R., Aunola, K., Leskinen, E., Nurmi, J., & Kirby, J. R. (2005). Development of individual differences in reading: Results from longitudinal studies in English and Finnish. *Journal of Educational Psychology*, 97, 299–319.
- Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University Press.
- Perfetti, C. A. (1992). The representation problem in reading acquisition. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 145–174). Hillsdale, NJ: Erlbaum.
- Schmidtke, A., Schaller, S., & Becker, P. (1978). *Raven-Matrizen Test: Coloured progressive matrices*. Weinheim, Germany: Beltz.
- Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, 55, 151–218.
- Seymour, P. H. K., Aro, M., & Erskine, J. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, 94, 143–174.
- Snowling, M. (2000). *Dyslexia* (2nd ed.). Oxford, England: Blackwell.
- Thaler, V., Ebner, E. M., Wimmer, H., & Landerl, K. (2004). Training reading fluency in dysfluent readers with high reading accuracy: Word specific effects but low transfer to untrained words. *Annals of Dyslexia*, 54, 89–113.
- Torgesen, J., Rashotte, C., & Alexander, A. W. (2001). Principles of fluency instruction in reading: Relationships with established empirical outcomes. In M. Wolf (Ed.), *Dyslexia, fluency, and the brain* (pp. 333–355). Timonium, MD: York Press.
- Torgesen, J., Wagner, R. K., Rashotte, C. A., Burgess, S., & Hecht, S. (1997). Contributions of phonological awareness and rapid automatic naming ability to the growth of word-reading skills in second- to fifth-grade children. *Scientific Studies of Reading*, 1, 161–185.
- Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of Child Psychology and Psychiatry*, 45, 2–40.
- Walczyk, J. J., Marsiglia, C. S., Johns, A. K., & Bryan, K. S. (2004).

- Children's compensations for poorly automated reading skills. *Discourse Processes*, 37, 47–66.
- Wimmer, H. (1993). Characteristics of developmental dyslexia in a regular writing system. *Applied Psycholinguistics*, 14, 1–33.
- Wimmer, H., & Hummer, P. (1990). How German-speaking first graders read and spell: Doubts on the importance of the logographic stage. *Applied Psycholinguistics*, 11, 349–368.
- Wimmer, H., & Landerl, K. (1997). How learning to spell German differs from learning to spell English. In C. A. Perfetti, L. Rieben, & M. Fayol (Eds.), *Research, theory, and practice across languages* (pp. 81–96). Mahwah, NJ: Erlbaum.
- Wimmer, H., Landerl, K., Linortner, R., & Hummer, P. (1991). The relationship of phonemic awareness to reading acquisition: More consequence than precondition but still important. *Cognition*, 40, 219–249.
- Wimmer, H., & Mayringer, H. (2002). Dysfluent reading in the absence of spelling difficulties: A specific disability in regular orthographies. *Journal of Educational Psychology*, 94, 272–277.
- Wimmer, H., Mayringer, H., & Landerl, K. (2000). The double-deficit hypothesis and difficulties in learning to read a regular orthography. *Journal of Educational Psychology*, 92, 668–680.
- Zoccolotti, P., De Luca, M., Di Pace, E., Gasperini, F., Judica, A., & Spinelli, D. (2005). Word length effect in early reading and in developmental dyslexia. *Brain and Language*, 93, 369–373.

Received May 23, 2006

Revision received July 26, 2007

Accepted August 21, 2007 ■

## Low Publication Prices for APA Members and Affiliates

**Keeping you up-to-date.** All APA Fellows, Members, Associates, and Student Affiliates receive—as part of their annual dues—subscriptions to the *American Psychologist* and *APA Monitor*. High School Teacher and International Affiliates receive subscriptions to the *APA Monitor*, and they may subscribe to the *American Psychologist* at a significantly reduced rate. In addition, all Members and Student Affiliates are eligible for savings of up to 60% (plus a journal credit) on all other APA journals, as well as significant discounts on subscriptions from cooperating societies and publishers (e.g., the American Association for Counseling and Development, Academic Press, and Human Sciences Press).

**Essential resources.** APA members and affiliates receive special rates for purchases of APA books, including the *Publication Manual of the American Psychological Association*, and on dozens of new topical books each year.

**Other benefits of membership.** Membership in APA also provides eligibility for competitive insurance plans, continuing education programs, reduced APA convention fees, and specialty divisions.

**More information.** Write to American Psychological Association, Membership Services, 750 First Street, NE, Washington, DC 20002-4242.



# Early First-Language Reading and Spelling Skills Predict Later Second-Language Reading and Spelling Skills

Richard L. Sparks  
College of Mount St. Joseph

Jon Patton, Leonore Ganschow, and  
Nancy Humbach  
Miami University

James Javorsky  
Oakland University

This prospective study examined early first-language (L1) predictors of later second-language (L2) reading (word decoding, comprehension) and spelling skills by conducting a series of multiple regressions. Measures of L1 word decoding, spelling, reading comprehension, phonological awareness, receptive vocabulary, and listening comprehension administered in the 1st through 5th grades were used as predictors of L2 reading (word decoding, comprehension) and spelling skills in high school. The best predictor of L2 decoding skill was a measure of L1 decoding, and the best predictors of L2 spelling were L1 spelling and L1 phonological awareness. The best predictor of L2 reading comprehension was a measure of L1 reading comprehension. When L2 word decoding skill replaced L1 word decoding as a predictor variable for L2 reading comprehension, results showed that L2 word decoding was an important predictor of L2 reading comprehension. The findings suggest that even several years after students learn to read and spell their L1, word decoding, spelling, and reading comprehension skills transfer from L1 to L2.

*Keywords:* L1, L2, reading, spelling, cross-linguistic transfer

Until recently, there was limited interest by foreign (second) language educators and researchers in examining the relationship between first, or native, language (henceforth, L1) skills and the next language to be learned, be it a second language learned in a natural setting or a foreign language learned in a classroom (henceforth, L2). Recently, however, with the proliferation of worldwide migration and the growing awareness of its implications for multilingual societies, researchers have begun to pay attention to L1–L2 connections. In the past few years, there has been an explosion in the number of studies investigating this relationship. In particular, there has been an increased interest in the extent to which L1 factors influence L2 reading development (e.g., see reviews by Geva & Verhoeven, 2000; Koda, 2005) and L2 oral and written proficiency (e.g., see review by Ganschow & Sparks, 2001). However, only a few of these studies have followed students' language performance from their primary school years into high school, at which time students in the United States begin to study an L2.

In this study, we followed students over a 10-year period to determine whether performance on measures of L1 reading (decoding, comprehension), spelling, phonological awareness, vocabulary, and listening comprehension are predictive of L2 reading (word decoding, comprehension) and spelling. Examination of the predictive value of L1 skills in elementary school for L2 skills in high school over a lengthy time period might provide useful information to researchers interested in issues of long-term cross-linguistic transfer of reading and spelling skills with older students learning an L2 years after they had acquired L1 reading and spelling skills. From a practical perspective, there may also be issues of interest for the prediction of L2 learning under different social and motivational contexts, for example, additive versus subtractive bilingualism, and for the diagnosis and possible prevention of reading and spelling problems in the L2.

## L1–L2 Relationships

Researchers have speculated that there is a relationship between students' L1 and L2 skills. Cummins (1979, 1984) developed the linguistic interdependence hypothesis, in which he argued that language and literacy skills can be transferred from one language to another. His primary contention was that success in the L2, for example, reading, depended on previous competence in L1 literacy skills. Researchers investigating Cummins's (1979, 1984) hypothesis have provided evidence of the interdependence of L1 and L2 skills among school-age English learners (e.g., see Cummins & Mulchay, 1978; Legarretta, 1979). In their linguistic coding differences hypothesis, Ganschow and Sparks (2001) proposed that

---

Richard L. Sparks, Department of Education, College of Mount St. Joseph; Jon Patton, Information Technology Services, Miami University; Leonore Ganschow, Department of Educational Psychology, Miami University; Nancy Humbach, Department of Teacher Education, Miami University; James Javorsky, Department of Human Development and Child Studies, Oakland University.

Correspondence concerning this article should be addressed to Richard L. Sparks, Department of Education, College of Mount St. Joseph, 5701 Delhi Road, Cincinnati, OH 45233. E-mail: richard\_sparks@mail.msjeu

L1 skills serve as the foundation for L2 learning and that problems with one component of language (e.g., phonological processing) are likely to have a negative effect on both L1 and L2 learning (e.g., see Sparks, 1995; Sparks & Ganschow, 1993). Their studies have shown that L2 learners in high school who achieve higher levels of oral (listening, speaking) and written (reading, writing) L2 proficiency and classroom achievement (grades) have significantly stronger L1 ability, especially phonological processing skills, than do L2 learners with lower levels of L2 proficiency and achievement (e.g., see Sparks et al., 1998; Sparks, Ganschow, Artzer, Siebenhar, & Plageman, 1998). In a recent study, Sparks, Humbach, and Javorsky (in press) found that high school students who had significant differences in L2 oral and written proficiency, word decoding, and spelling exhibited significant differences in their L1 literacy skills as early as the fourth grade. In another study, Sparks, Patton, Ganschow, Humbach, and Javorsky (2006) found that L1 reading and spelling skills in elementary school were the best predictors of oral and written L2 proficiency in high school.

Studies by other researchers have also found a strong relationship between L1 skills and L2 proficiency in a variety of languages. Dufva and Voeten (1999) tested 160 Finnish students at the beginning of the first grade and followed them into the third grade, when they started to learn English. They verified that both L1 literacy and phonological memory skills were predictive of L2 learning. Olshtain, Shohamy, Kemp, and Chatow (1990) discovered that L1 academic proficiency played the most important role in L2 learning among a group of Hebrew-speaking students learning English. In a recent study with U.S. college students, Meschyan and Hernandez (2002) found that L1 decoding skills predicted college-age adults' L2 competency and that this relationship was mediated by participants' L2 word decoding skills. They also found that L1 decoding skill was an important predictor of grades in introductory L2 courses. Other investigators have obtained findings similar to the aforementioned studies (e.g., Holm & Dodd, 1996; Hulstijn & Bossers, 1992; Humes-Bartlo, 1989; Kahn-Horwitz, Shimron, & Sparks, 2005, 2006).

### L1-L2 Decoding

Related research on the processes by which students learn to read their L1 provides potential insights that assist our understanding about the role of decoding in learning to read an L2. There is considerable research evidence indicating that students who become skilled readers are those who read (decode) words accurately and fluently (see reviews by Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001; Wolf, 2001). Learning to read words requires knowledge of phonological awareness, letter-sound correspondences, and orthographic patterns (Snow, Burns, & Griffin, 1998). In turn, accurate word decoding skills enable students to read most words in the text to comprehend its meaning (Ehri, 2005). Longitudinal research has shown that the ability to read words in the first grade predicts reading-related skills 11 years later (Cunningham & Stanovich, 1997; MacDonald & Cornwall, 1995).

In contrast to the voluminous research conducted on L1 word decoding, Koda (2005) has reported that L2 word decoding has received little attention from researchers for two reasons.<sup>1</sup> First, top-down conceptualizations of L2 reading (e.g., see Bernhardt, 1991) have dominated L2 research in the past. Second, linguistic

knowledge has often been equated with language processing skills (e.g., see Bialystock, 2001). Thus, reading skills, and specifically word decoding, were thought to develop concurrently with increased knowledge of the L2. However, little or no theoretical support has been found for this hypothesis.

Recent studies have found that L1 phonological awareness is a good predictor of subsequent reading and decoding skills in another language, that is, cross-linguistic transfer (Durgunoglu, 2002). Investigations of cross-linguistic transfer in the development of literacy skills have shown that phonological awareness skills are correlated across languages, particularly with word recognition skills across languages (e.g., see Cisero & Royer, 1995; Geva, Yaghoub-Zadeh, & Schuster, 2000). For example, examining English-language learners whose L1 was Spanish, Lindsey, Manis, and Bailey (2003) found that phonological awareness skills transferred from Spanish to English and predicted word recognition skills in English. In another study with English-speaking children in French immersion classes, Comeau, Cormier, Grandmaison, and Lacroix (1999) confirmed that phonological awareness in both languages was associated with increases in skill at decoding French words.

L2 researchers have also conducted studies comparing high- and low-proficiency L2 readers on word decoding tasks. In one set of studies, low-proficiency readers exhibited slow and less accurate performance on word decoding tasks (e.g., see Favreau & Segalowitz, 1982; Haynes & Carr, 1990). In another group of studies, low-proficiency L2 readers relied less on discourse-level and more on word-level processing (Cziko, 1980; Horiba, 1990) and on a word's graphic cues rather than on semantic information (Chamot & El-Dinary, 1999; Clarke, 1980). The findings suggest that similar to poor L1 readers, poor L2 readers with weak word decoding skills may spend more time sampling visual information than using multiple information sources (Koda, 2005). In other studies, Segalowitz and his colleagues verified that fluent bilingual speakers displayed large variations in their word decoding fluency (Favreau & Segalowitz, 1983; Segalowitz, 1986; Segalowitz, Poulson, & Komoda, 1991). Such findings imply that word decoding skills develop independently of oral language competence, and the acquisition of oral language in the L2 does not ensure development of word recognition skills. Moreover, longitudinal studies with school-age L2 learners have revealed no systematic link between L2 word reading skills and L2 oral language proficiency (August, Calderon, & Carlo, 2001; Gholamain & Geva, 1999).

Recent studies have also shown that the orthographic depth of the L2 (i.e., orthographies with highly regular sound-symbol relationships vs. orthographies that preserve morphological information at the expense of phonological transparency), the orthographic distance between the L1 and L2 (i.e., similarities and dissimilarities in L1-L2 writing systems), and L1 experience and knowledge contribute to skill in L2 word decoding. In their orthographic depth hypothesis, Katz and Frost (1992) proposed that the orthographic depth, that is, shallow versus deep orthographies, of a written language affects word decoding skill. Shallow (transparent) or-

<sup>1</sup> Koda (2005, p. 29) distinguished between *word recognition* (i.e., "the process of obtaining words' sounds and meanings") and *decoding* (i.e., "extraction of phonological information"). In this article, the two terms are used interchangeably.



thographies map consistently the sounds of a language with a specific symbol; however, deep (opaque) orthographies have more variable sound–symbol correspondences and can map a sound with more than one symbol. Cross-linguistic studies with L2 learners have found that they achieve faster and more accurate L2 word recognition skills if the L2 orthography is more closely related to their L1 orthography (e.g., see Akamatsu, 1999; Koda, 1999; Muljani, Koda, & Moates, 1998). Other studies have found that L2 learners use identifiable L1 processing skills for L2 word decoding (e.g., see Brown & Haynes, 1985; Ryan & Meara, 1991). In one study, Koda (1990) discovered that learners of English as a second language (ESL) were able to read phonetically accessible nonsense words but not phonetically inaccessible (Sanskrit) ones. Her findings corroborate research that has found that L2 learners with different orthographic backgrounds use distinct processing skills to read words. She concluded that L1 orthographic experience has an impact on L2 lexical processing and that L1 processing is a major source of variance among L2 learners' word decoding skills (see Koda, 2005).

### L1–L2 Spelling

Research on the processes by which students learn to spell their L1 also provides potential insights about the role of L1 spelling skills in learning to spell an L2. In the L1, researchers have revealed that phonological awareness skills are highly correlated with spelling ability (Bruck & Treiman, 1990; Stuart & Masterson, 1992) and that children's ability to spell is also related to knowledge of phonological processing and letter–sound correspondences in that sounds map onto letters (e.g., see Moats, 2005; Treiman, 1993). Frith (1980) proposed that spelling is “visible phonology” (p. 2) and that it also reflects other levels of language, chiefly visual or orthographic factors.

Investigations of the development of L2 spelling skills have examined how phonological processing in the L1 affects spelling skills in the L2. In a study with ESL learners, Wade-Woolley and Siegel (1997) showed that spelling skills in the L2 were associated with phonological processing skills, but students' reading skill in the L2 had a more significant influence on spelling accuracy than their L1 skills. In a study with L1 and minority (immigrant) children learning to read and spell in an L2 in the Netherlands, Verhoeven (2000) found that although the minority children were less efficient in various reading and spelling processes in Dutch than were their native Dutch peers, L1 and L2 readers and spellers relied on highly comparable strategies. In addition, orthographic factors also are likely to influence spelling performance. For example, Treiman (1993) has shown that children are sensitive at an early age to the orthographic structures in the spellings of words. In a review of several studies on basic processes in early L2 reading and spelling, Geva and Verhoeven (2000) concluded that, “L1–L2 transfer at the level of underlying component skills such as phonological awareness relates to word recognition and spelling” (p. 265).

### L2 Reading Comprehension

Although new findings in L2 reading research have shown that readers must decode words efficiently to comprehend text in the L2, there is still debate about the precise contributions of lower

order or bottom-up and higher order top-down processes for L2 reading comprehension (Koda, 2005). In L1 research, Cutting and Scarborough (2006) reviewed evidence that showed that both bottom-up, that is, accurate word decoding, and top-down, that is, proficient use of semantic and syntactic relationships for understanding the meaning of text, are necessary for skilled reading comprehension. Koda (1992) has noted that most investigations of L2 reading have focused on higher order processes, such as linguistic knowledge, background knowledge, and cognitive and metacognitive skills, in explaining reading comprehension skills. In her reviews of the L1 and L2 research literature, however, she found that inefficient and slow word decoding in the L2 plays a significant role in L2 reading comprehension (Koda, 1998) and concluded that decoding skills are a “precondition” for L2 comprehension (Koda, 2005, p. 256). In recent years, some L2 studies have reported positive correlations between lower order (decoding) skills and reading comprehension (Hirai, 1994; Nassaji & Geva, 1999). In one study, VanGeldereren et al. (2004) found that although the componential structures of L1 and L2 reading comprehension are not identical, L1 reading comprehension made a large contribution to L2 reading comprehension. Fukkink, Hulstijn, and Sims (2005) found that although L2 word decoding is important for L2 reading comprehension, there may not be a strong causal link between an “increase of lexical access, beyond a certain threshold that has yet to be determined, and reading comprehension” (p. 72). They concluded that automatic lexical access helps in processing words in the text, but there are many processes operating simultaneously that contribute to efficient L2 reading comprehension. In a study of fourth-grade learners of English as a foreign language, Proctor, Carlo, August, and Snow (2005) showed that both L2 oral language skills (vocabulary knowledge, listening comprehension) and L2 word decoding skills (alphabetic knowledge, fluency) are important for L2 reading comprehension.

### Summary and Hypotheses

Evidence indicates that students' facility with reading and spelling in the L1 is related to their learning to read and spell in L2, demonstrating cross-linguistic transfer of skills. However, the assertion that L1 reading and spelling skills in elementary school have an effect on L2 reading and spelling skills in high school several years later would entail long-term cross-linguistic transfer of L1 skills to the L2. Koda (2005) described the impact that connectionist views of reading acquisition have had on research and thinking about the development of L2 reading skills. In the connectionist model, units are linked to one another to form a network. Retrieval of specific activation patterns stored in memory (i.e., letter strings, letter–sound correspondences) is effortless and is likely to be activated by L2 input, regardless of the orthographic distance between L1 and L2. In her view, the connectionist model of automaticity establishes a rationale for predicting “procedural variations stemming from L1 skill activation among L2 readers with diverse backgrounds” (Koda, 2005, p. 18). The connectionist model offers a way to study and analyze L2 reading competence and its development because the ability to read an L2 requires the ability to “map between language forms and their functions. . . with increased experience in the L2, form–function mapping procedures are activated automatically irrespective of the learner's intention” (Koda, 2005, p. 18).

Building on the connectionist framework, we propose that because the fundamental competencies—decoding and comprehension—are the same in L1 and L2 reading, there is reason to expect that there will be long-term cross-linguistic transfer of L1 reading and spelling skills developed in elementary school to L2 reading and spelling skills several years later when learners first encounter an L2 in high school. Specifically, students' decoding and spelling skills in their alphabetic L1 (English) will likely account for a large part of the variance in their decoding and spelling skills in an alphabetic L2 (Spanish, French, German). Likewise, L1 decoding and L1 reading comprehension skills will likely account for part of the variance in L2 reading comprehension skills. However, as students' exposure to the L2 increases over 2 years of L2 study in high school, their L2 decoding skill may account for a larger part of the variance in L2 reading comprehension than either their L1 decoding or L1 reading comprehension skills.

The present study is one part of a larger prospective investigation conducted with the participants. In a previously published study Sparks et al. (2006) examined the best predictors of overall L2 proficiency (i.e., listening comprehension + speaking + writing + reading comprehension) and of L2 aptitude measured by a standardized L2 aptitude test, the Modern Language Aptitude Test (Carroll & Sapon, 1959). The study reported here differs from the aforementioned study by specifically examining the role of L1 skills, that is, word decoding, reading comprehension, spelling, receptive vocabulary, and listening comprehension, in predicting students' L2 word decoding, reading comprehension, and spelling skills rather than their overall oral and written L2 proficiency. It also differs by investigating the long-term prediction of L2 reading and spelling skills.

## Method

### Participants

The participants were 54 high school students, 25 male and 29 female, from a large middle-class rural school district in the Midwest. All of the students had completed 2 years of L2 study in one of three languages—Spanish ( $n = 30$ ), French ( $n = 14$ ), and German ( $n = 10$ ). The students were followed from the 1st through the 10th grades. All of the participants completed their second year of the L2 by the end of the 10th grade. The mean age of the participants at the beginning of the 1st grade was 6 years, 9 months; at the end of the study 10 years later, the mean age of the participants was 16 years, 4 months (age range = 15 years, 9 months to 16 years, 11 months).<sup>2</sup> All of the students were Caucasian. The study began with 156 students when they entered 1st grade in this school district. By the 9th grade, 101 of the 156 students still attended school in the district. Seventy-seven of the 101 students began to learn an L2 in the 9th grade. Fifty-four of the 77 students chose to continue their participation in the study when they began their first year in an L2 course. Parental permission was obtained for each participant.

Resources were not available to test all 1st-grade students enrolled in the district's elementary school in a given year. However, we wanted to obtain a sample of students who would be likely to take an L2 course in high school and to exhibit a range of academic achievement skills. Therefore, we chose a cohort model and used a sample of good, average, and below average readers over a

3-year period. The 54 participants were chosen in the following manner. The school district administered a reading readiness measure, the Test of Reading Readiness: Level K (Scott, Foresman and Company, 1987) to all of its kindergarten students each year. On the basis of their scores on this measure (raw scores ranged from 0 to 34) and the kindergarten teacher's recommendation, students were assigned by the school to one of three types of 1st-grade classrooms: (a) above average readers, (b) at-risk readers, and (c) average readers. In the school, there was one class for above average readers, one class for at-risk readers, and several classes for average readers each year. For each of the three cohorts, we invited all of the students in the above average class and the at-risk class to participate in the study. On the Test of Reading Readiness: Level K, the above average students had achieved raw scores ranging from 33 to 34 and the at-risk students had achieved raw scores ranging from 0 to 16. A random sample of students in the average reading classes whose score on the reading readiness measure ranged from 17 to 33 were selected to take part in the study. Of the 54 students who participated for all 10 years of the study, 18 (33.3%) were above average readers, 19 (35.2%) were average readers, and 17 (31.5%) were at-risk readers. There were similar numbers of students in each of the three cohorts (Cohort 1 = 17 students, Cohort 2 = 20 students, Cohort 3 = 17 students).

## Instruments

### Measures of L1 Proficiency

**Word decoding.** The measure of L1 word decoding was the Woodcock Reading Mastery Test—Revised Basic Skills Cluster, Forms G and H (Woodcock, 1987). The Basic Skills Cluster is composed of two subtests, Word Identification and Word Attack. On the Word Identification subtest, a student read aloud a list of increasingly difficult words. On the Word Attack subtest, a student read aloud a list of increasingly difficult pseudowords. For a response to be considered correct, the student had to produce a natural reading (pronunciation) of the word or pseudoword. A test-retest reliability of .96 was reported for the Basic Skills Cluster.

**Spelling.** The measure of L1 spelling was the Test of Written Spelling—2 (Larsen & Hammill, 1986). On this dictated word test, the student wrote the words spoken by the examiner. The response was marked as correct or incorrect. A test-retest reliability of .95 was reported for this test.

**Reading comprehension.** The measure of L1 reading comprehension was the Formal Reading Inventory, Forms A and B (Wiederholt, 1986). On this multiple-choice test, the student answered five questions after reading silently a short paragraph. The response was marked as correct or incorrect. Internal consistencies of .92 to .97 were reported for the two forms of the test used in this study.

**Phonological awareness.** The measure of L1 phonological awareness was the Lindamood Auditory Conceptualization Test, Forms A and B (Lindamood & Lindamood, 1979). On this test, the student manipulated blocks of various colors to indicate his or her

<sup>2</sup> We extend special thanks to Jackie Blair for her crucial role as facilitator and school liaison in this study. The study could not have been completed without her assistance.



conceptualization of speech sounds spoken by the examiner. The student showed the number, sameness, or difference of sounds (e.g., show me /b/ /b/ /z/) and represented with the blocks changes in spoken patterns without associating the sounds with letters (e.g., if that says *ups*, show me *usp*). This test yielded raw scores with a range of 0 to 100. Pretest–posttest reliability of .96 was reported for the two forms of the test.

*Reading readiness.* The measure of reading readiness was the Test of Reading Readiness: Level K. The test consists of two parts, each of which had several subtests. In this study, Part 2, Level B was used to gauge reading readiness. Level B consisted of four subtests: the Auditory Discrimination/Rhyme subtest tested the ability to listen to a spoken word (“dish”) and match a picture that rhymed (*fish*); the Letter/Sound Relationships subtest tested the ability to look at a letter (*w*), know its sound (/w/), and match the sound with a picture that started with the same sound (*worm*); the Word Recognition subtest tested the ability to listen to three words (“dog,” “does,” “blue”) and to circle the written word with a different initial sound than the others (*blue*); and the Sentence Reading subtest tested skill for choosing the correct written sentence (out of two sentences) that matched the action in an accompanying picture. The four subtests produced a raw score with a range of 0 to 37. A test–retest reliability of .90 was reported.

*Vocabulary.* To assess L1 vocabulary, the Peabody Picture Vocabulary Test—Revised, Forms L and M (Dunn & Dunn, 1981) was used. The test measured receptive vocabulary for standard American English. On this test, the student was shown four pictures and asked to identify the picture for the word spoken by the examiner. The response was marked as correct or incorrect. A median test–retest reliability of .82 was reported for the two forms of the test.

*Listening comprehension.* The Woodcock Reading Mastery Test—Revised Passage Comprehension subtest, Forms G and H (Woodcock, 1987) was used to appraise L1 listening comprehension. This cloze test consists of reading a short passage (one–two sentences) aloud to a student and asking him or her to identify aloud a key word missing from the passage. The student was not permitted to read (see) the passage, but the passage could be repeated. Again, the response was marked as correct or incorrect. This subtest is generally used as a measure of reading comprehension; however, the aforementioned alternative procedure was recommended by Aaron (1989) as a diagnostic indicator in identifying problem readers and was used for this study. A test–retest reliability of .92 when used as a measure of reading comprehension was reported for the two forms of this subtest.

Each of the preceding L1 achievement measures was chosen to examine whether the processes used in learning to read and spell the L1 would be important for learning to read (decode, comprehend) and spell the L2 several years later. The L1 phonological awareness, reading readiness, and word decoding measures were expected to correlate with a measure of L2 word decoding. The L1 phonological awareness and spelling measures were expected to correlate with a measure of L2 spelling. The L1 reading comprehension, listening comprehension, and vocabulary measures were predicted to correlate with a measure of L2 reading comprehension.

## *Measures of L2 Word Decoding, Spelling, and Reading Comprehension*

Nonstandardized L2 tasks of word decoding, spelling, and reading comprehension in Spanish, French, and German were designed by three L2 educators who had expertise in their language of study. L2 proficiency measures are not standardized and do not yield standardized scores for comparison purposes. Instead, the proficiency guidelines developed by the American Council for the Teaching of Foreign Languages (ACTFL; ACTFL, 1989) list levels of proficiency that range from novice–low to superior. The guidelines do not distinguish between oral versus written expression or reading versus writing and do not include word decoding and spelling as separate proficiency skills.

*L2 word decoding.* The word decoding task for each of the three languages consisted of a list of 20 actual words and a list of 20 pseudowords. The real and pseudowords in the three L2s measured a specific decoding task that involved use of vowel sounds in each L2 that were not consistent with the vowel sounds in English. The vowel sounds in the L2s used different letter combinations and contained multisyllabic words and words with diacritical markings. Each student read aloud the list of real words and pseudowords. For a response to be considered correct, the student had to pronounce the word correctly. If the student spontaneously self-corrected a response, then that response was counted as correct. The reliability of the L2 word decoding lists and pseudoword lists was checked by calculating Cronbach’s alpha. For the Spanish words, the reliability coefficient for the word decoding list was .75 and for the pseudoword decoding list was .75. For the French words, the reliability coefficient for the word decoding list was .78 and for the pseudoword decoding list was .84. For the German words, the reliability coefficient for the word decoding list was .81 and for the pseudoword decoding list was .79. The word lists in the three L2s are presented in Appendix A.

*L2 spelling.* The L2 spelling task for each of the three languages consisted of 20 words designed to measure a specific decoding task. Like the word decoding tasks, this measure included vowel sounds in the L2 that were not consistent with the vowel sounds in English, used different letter combinations, and contained multisyllabic words and words with diacritical markings. The students wrote the words spoken by the examiner, and their responses were marked as correct or incorrect. The reliability of the L2 spelling lists was checked by calculating Cronbach’s alpha. For the Spanish words, the reliability coefficient was .78, for the French words, the reliability coefficient was .81, and for the German words, the reliability coefficient was .74. The spelling lists are presented in Appendix B.

*L2 reading comprehension.* The L2 reading comprehension measure tested students’ comprehension using the ACTFL proficiency guidelines (ACTFL, 1986, 1989). The L2 reading comprehension test was designed using criteria descriptive of the intermediate–high level of the ACTFL guidelines. The reading comprehension test in Spanish used the same items and prompts as the French and German tests. Students read silently a one-page letter written in the L2 and answered 10 multiple-choice questions in English about the letter by circling the correct answer. The students were then given a passage to read silently from *Reader’s Digest* in Spanish, French, or German and were asked to answer 10 multiple-choice questions in English about the passage by circling

the correct answer. The only differences in the three reading comprehension tests were those specific to a particular L2. The test directions and time limits were the same for each of the three L2s. The comprehension questions were not ordered by degree of difficulty. The Cronbach's alpha was .52.

### Procedure

Four of the L1 measures (i.e., Woodcock Basic Skills Cluster, Test of Written Spelling—2, Formal Reading Inventory, and Peabody Picture Vocabulary Test—Revised) were administered in elementary school at five different time intervals: at the beginning of the first grade and the end of first, second, third, and fifth grades. The Lindamood Auditory Conceptualization Test was administered only through the third grade, and the listening comprehension test (Woodcock Passage Comprehension subtest) was administered only at the end of the third and fifth grades. For each of the L1 measures, a student's scores were combined across years (i.e., first through fifth grades) to obtain a mean score on each measure. The L1 measures were administered by Richard L. Sparks and Leonore Ganschow, with assistance from undergraduate and graduate students trained prior to administration of the tests.<sup>3</sup> Scores on the reading readiness measure, the Test of Reading Readiness: Level K, were obtained from school records at the end of the first and second grades.

The L2 word decoding measure was administered individually to each student at the end of the 9th- and 10th-grade years by Nancy Humbach and graduate students trained by her.<sup>4</sup> For the L2 word decoding measure, the student read the word lists twice, once at the end of the first-year L2 course and again at the end of the second-year L2 course. The maximum score for the first-year (9th grade) L2 word decoding measure was 40 words (20 real words and 20 pseudowords); likewise, the maximum score for the second-year (10th grade) L2 word decoding measure was 40 words. The students' performance for each of the 2 years was analyzed separately because we speculated that their performance on the word lists might increase as they became more familiar with the L2 being studied.

The L2 spelling measure was given individually to each student at the end of the 10th-grade year by Nancy Humbach and graduate students trained by her. The student wrote each word on a piece of paper as the examiner read the word aloud. The examiner repeated a word as often as necessary. The maximum score on the spelling test was 20 words.

The L2 reading comprehension measure was administered in groups by Richard L. Sparks at the end of the students' second-year L2 course. The students were given 15 min to read the letter and answer the questions and 15 min to read the *Readers Digest* passage and answer the questions. The maximum score for the assessment was 20.

Analyses of variance were used to determine whether there were significant differences among the three different L2 groups (Spanish, French, and German) on the four L2 measures. Findings showed that there were no significant differences among the three L2 groups on the L2 word decoding test, Year 1,  $F(2, 51) = 0.05$ ,  $p = .95$ ; the L2 word decoding test, Year 2,  $F(2, 51) = 0.73$ ,  $p = .49$ ; the L2 spelling test,  $F(2, 51) = 1.25$ ,  $p = .30$ ; and the L2 reading comprehension test,  $F(2, 51) = 1.34$ ,  $p = .27$ .

### Results

Table 1 reports the mean scores of the 54 participants on the L1 variables. For the L2 measures, the mean score on the L2 word decoding measure at the end of the first-year L2 course was 20.7 ( $SD = 6.4$ ); at the end of the second-year L2 course, the students' mean score on the L2 word decoding measure was 25.7 ( $SD = 7.1$ ). The mean score of the 54 students on the L2 spelling measure at the end of the second-year L2 course was 8.8 ( $SD = 3.6$ ). The mean score of the 54 students on the L2 reading comprehension measure at the end of the second-year L2 course was 10.2 ( $SD = 3.0$ ). Table 2 presents intercorrelations among the L1 and L2 measures for the total group.

Four separate analyses were conducted to determine which L1 variables in elementary school were the best predictors of L2 word decoding, Year 1; L2 word decoding, Year 2; L2 spelling; and L2 reading comprehension in high school. Two additional analyses were then conducted in which the L2 decoding measures replaced the L1 decoding test to predict L2 reading comprehension. In the first analysis, the L2 word decoding, Year 1 measure replaced the L1 decoding measure, Woodcock Basic Skills Cluster, as a predictor variable for L2 reading comprehension. In the second analysis, the L2 word decoding, Year 2 test replaced L2 word decoding, Year 1 (and Woodcock Basic Skills Cluster) as a predictor variable for L2 reading comprehension.

Multiple regression analyses using the maximum  $R^2$  improvement (MAXR) technique for variable selection was applied to determine which variables were important predictors of L2 word decoding, L2 spelling, and L2 reading comprehension. This procedure is sometimes used on data sets of smaller size so that relationships of interest that may exist in the population might not be overlooked. The MAXR first calculates the best one-variable model (i.e., which single measure best correlates with a specific outcome variable). Then, the best two-variable model is selected on the basis of the maximum  $R^2$  (coefficient of determination) improvement. The best two-variable model indicates the two measures that explain the highest percentage of the variance in L2 proficiency about its mean. This process continues by determining the best three-variable model yielding the largest  $R^2$  improvement and so on, until a model containing all the measures as predictors of an outcome variable is generated. Unlike the traditional stepwise selection methods (e.g., forward selection, backward elimination, stepwise regression) that provide one best variable model, the MAXR procedure provides several alternative models from which to choose and does not perform any inferential tests in selecting its best models. To ascertain the size of the best model to select, Mallow's  $C_p$  statistic was used. The model having a  $C_p$  statistic that is less than the number of predictors allows the choice of a model having a small mean square error and small regression bias. The mean square error gauges the average squared deviation between the actual score on an outcome variable for a student and the score predicted by the regression procedures, taking into account the degrees of freedom in the data. A model with minimal

<sup>3</sup> We extend special thanks to Chris Utter and her staff for their very important roles in the completion of this study.

<sup>4</sup> We thank Kelly Noe and Johannes Tokarski for administering the (real) word decoding, pseudoword decoding, and spelling measures to Cohorts 2 and 3.



Table 1  
*Means (and Standard Deviations) of Participants on the First Language (L1) Measures*

Measure	<i>M (SD)</i>	Minimum	Maximum
Woodcock Basic Skills Cluster <sup>a</sup>	102.0 (12.5)	68.4	131.6
Test of Written Spelling—2 <sup>a</sup>	95.9 (11.6)	75.2	124.2
Formal Reading Inventory <sup>a</sup>	101.9 (11.1)	80.8	131.4
Peabody Picture Vocabulary Test—Revised <sup>a</sup>	106.7 (11.2)	84.0	136.2
Lindamood Auditory Conceptualization Test <sup>b</sup>	0.0 (1.0)	−1.93	1.62
Test of Reading Readiness <sup>b</sup>	0.0 (1.0)	−2.94	0.93
Listening comprehension <sup>a</sup>	103.9 (10.4)	83.0	124.0

<sup>a</sup> Standard scores, *M* = 100, *SD* = 15. <sup>b</sup> *z* scores.

regression bias has strong predictive power on observations not necessarily included in the sample from which the model is based. The final model chosen is the one having all of the variables that show mild or strong significance and are not suppressors.

For the prediction models, a tolerance diagnostic was computed for each predictor variable to determine whether multicollinearity among the variables might exist. A tolerance of less than .10 would indicate the presence of significant multicollinearity problems (Mendenhall & Sincich, 1989). The tolerance was at or greater than .26 (range = .26–.70).

*L2 Word Decoding, Year 1*

The analysis yielded a one-variable model in which the best predictor was the Woodcock Basic Skills Cluster ( $\beta = .72$ ),  $F(1, 52) = 57.16$ ,  $p < .0001$ . The Woodcock Basic Skills Cluster alone accounted for 52% of the variance.

*L2 Word Decoding, Year 2*

The analysis produced a one-variable model in which the best predictor was the Woodcock Basic Skills Cluster ( $\beta = .65$ ),  $F(1, 52) = 38.71$ ,  $p < .0001$ . The Woodcock Basic Skills Cluster accounted for 43% of the variance in the model.

Table 2  
*Intercorrelations Among First Language (L1) Variables, Second Language (L2) Word Decoding, L2 Reading Comprehension, and L2 Spelling*

Measure	1	2	3	4	5	6	7	8	9	10	11
1. Basic Skills Cluster	—	.94**	.81**	.57**	.68**	.63**	.64**	.63**	.72**	.65**	.44**
2. Test of Written Spelling—2		—	.82**	.59**	.68**	.64**	.61**	.69**	.68**	.62**	.50**
3. Formal Reading Inventory			—	.73**	.59**	.66**	.73**	.62**	.59**	.59**	.50**
4. Peabody Picture Vocabulary Test—Revised				—	.46**	.48**	.75**	.45**	.50**	.37**	.32*
5. Lindamood Auditory Conceptualization Test					—	.52**	.49**	.66**	.59**	.48**	.44**
6. Test of Reading Readiness						—	.57**	.45**	.40**	.47**	.28*
7. L1 listening comprehension							—	.41**	.50**	.46**	.31*
8. L2 spelling								—	.57**	.66**	.56**
9. L2 word decoding, Year 1									—	.75**	.48**
10. L2 word decoding, Year 2										—	.62**
11. L2 reading comprehension											—

\*  $p < .05$ . \*\*  $p < .01$ .

*L2 Spelling*

The analysis generated a two-variable model in which the best predictors were the Test of Written Spelling—2 ( $\beta = .45$ ) and the Lindamood Auditory Conceptualization Test ( $\beta = .36$ ),  $F(2, 51) = 29.93$ ,  $p < .0001$ . The two variables accounted for 54% of the variance in the model. The Test of Written Spelling—2 alone accounted for 47% of the variance. Both variables were significant at the  $p < .01$  level.

*L2 Reading Comprehension (With Only L1 Measures as Predictor Variables)*

The analysis yielded a one-variable model in which the best predictor was the Formal Reading Inventory ( $\beta = .50$ ),  $F(1, 52) = 17.56$ ,  $p < .0001$ . The Formal Reading Inventory accounted for 25% of the variance.

*L2 Reading Comprehension (With L2 Word Decoding, Year 1 Replacing the Woodcock Basic Skills Cluster)*

The analysis yielded a two-variable model in which the best predictors were the Formal Reading Inventory ( $\beta = .34$ ) and L2 word decoding, Year 1 ( $\beta = .28$ ),  $F(2, 51) = 11.16$ ,  $p < .0001$ . The two variables accounted for 31% of the variance in the model. The Formal Reading Inventory alone accounted for 25% of the variance and was significant at the  $p < .05$  level. The L2 word decoding, Year 1 measure approached significance ( $p = .056$ ).

*L2 Reading Comprehension (With L2 Word Decoding, Year 2 Replacing L2 Word Decoding, Year 1 and the Woodcock Basic Skills Cluster)*

The analysis yielded a one-variable model in which the best predictor was L2 word decoding, Year 2 ( $\beta = .62$ ),  $F(1, 52) = 33.12$ ,  $p < .0001$ . The L2 word decoding, Year 2 measure alone accounted for 39% of the variance.

Discussion

This study examined the role of L1 skills (i.e., word decoding, phonological awareness, reading comprehension, spelling, recep-

tive vocabulary, and listening comprehension) in predicting L2 word decoding, spelling, and reading comprehension for high school students who had learned to read and spell their L1 several years earlier. In addition, the study examined whether L2 word decoding skill was important for predicting L2 reading comprehension. The results are discussed under the following headings: *L2 Word Decoding*, *L2 Spelling*, and *L2 Reading Comprehension*.

### *L2 Word Decoding*

The results of the regression analyses showed that the measure of L1 word decoding, the Woodcock Basic Skills Cluster, was the best predictor of L2 word decoding at the end of the first- and second-year L2 courses. In both years, L1 decoding skills at the end of the second, third, and fifth grades accounted for the largest part of the variance in L2 word decoding skills from several years later in high school. These findings are consistent with a growing body of research that has confirmed that the skills used to read words in the L1 are highly correlated with the skills used to read words in an L2 (e.g., see Geva, Wade-Woolley, & Shany, 1997; Kahn-Horwitz et al., 2006; Muljani et al., 1998). Moreover, the finding that students' performance on a measure of L1 decoding skill was related to L2 word decoding is consistent with studies that have revealed a relationship between L1 literacy skills and subsequent performance in L2 learning (e.g., see Dufva & Voeten, 1999; Geva et al., 1997; Meschyan & Hernandez, 2002; Sparks, Ganschow, et al., 1998) and supports the hypothesis that there is long-term cross-linguistic transfer of L1 phonological processing skills to L2 word decoding.

At the end of the first-year L2 course, L1 decoding skill accounted for over half (52%) of the variance in L2 word decoding; at the end of the second-year L2 course, L1 decoding skills accounted for just under half (43%) of the variance in L2 decoding skills. These observations give rise to an important question: Which factors might account for the remaining variance in L2 word decoding? Koda (2005) reported that processes other than phonological coding are used to decode L2 words. For example, she reviewed evidence that showed that a student's facility and subsequent experiences with his or her L1 may account for some of the variance in L2 decoding skills. She also noted that the orthographic distance between the L1 and L2 may account for some of the quantitative variations in L2 decoding skills (e.g., see Akamatsu, 1999; Muljani et al., 1998). A related question is why L1 decoding skills on the Woodcock Basic Skills Cluster explained more of the variance in first-year L2 word decoding than did second-year L2 word decoding skills, a finding that might be explained by the orthographic distance between the students' native language, English, and the three L2s they studied in high school. As the students received more exposure to written words in Spanish, French, or German from Year 1 to Year 2 of their high school L2 courses, they may have become more skillful in their use of strategies for decoding words in Spanish, French, or German. For example, students enrolled in Spanish may have become more accustomed to the regularity of the letter-sound correspondences for vowel sounds in Spanish and also to consonant sounds (e.g., *ll* = /y/) and vowel digraphs (e.g., *ua* = /wä/) that are unique to Spanish. With increased practice and exposure to the L2 orthography, they may have become more skillful in decoding words in the L2.

Another hypothesis for the aforementioned finding is that the three L2s in this study are more transparent orthographies than English. In their orthographic depth hypothesis, Katz and Frost (1992) proposed that phonological information is assembled through letter-by-letter translation in shallow orthographies but that in deep orthographies, phonological information is obtained after a word has been identified on the basis of stored knowledge of the word. In contrast to English, Spanish and German are shallow orthographies in which letters correspond to one phoneme. Although French is more opaque than Spanish and German, students in the three L2s may have learned the three orthographies well enough to rely less on their L1 experiences in English by the second year of the L2 course.

### *L2 Spelling*

The results indicated that the measure of L1 spelling, the Test of Written Spelling—2, was the best predictor of L2 spelling at the end of the second-year L2 course. Students' L1 spelling skills in elementary school explained 47% of the variance in L2 spelling in the 10th grade, which suggests that L1 spelling skills are used for learning to spell words in the L2 even after students learned to spell their L1 many years earlier. These findings are consistent with research that has shown that the component processes underlying L1 and L2 spelling are "highly comparable at the word level" (Verhoeven, 2000, p. 327).

In the prediction model, a measure of L1 phonological awareness in elementary school, the Lindamood Auditory Conceptualization Test, also contributed statistically to the variance in L2 spelling at the end of the second-year L2 course several years later. Phonological awareness (knowledge of the internal sound structure of spoken words) is one of the best predictors of learning to read, not only in English but also in several alphabetic languages (e.g., see Bradley & Bryant, 1983; Lundberg, Olofsson, & Wall, 1980; Rayner et al., 2001; Wagner & Torgesen, 1987). Likewise, phonological awareness skill is highly correlated with spelling ability (Bruck & Treiman, 1990; Stuart & Masterson, 1992). The fact that L1 phonological awareness in elementary school was predictive of L2 spelling ability several years later in high school suggests that long-term transfer of phonological awareness skill is important for spelling and decoding words in an alphabetic L2.

L1 spelling and phonological awareness skills accounted for just over 54% of the variance in L2 spelling skills. The findings for the prediction of L2 spelling are similar to those for L2 word decoding in this study; that is, skills related to phonological processing also accounted for approximately half of the variance in L2 word decoding. Like L2 word decoding, the additional variance in L2 spelling skills might be explained by the students' facility and experiences with their L1, which is consistent with recent evidence showing that L1 has "clearly detectable" impacts on L2 word processing (Koda, 2005, p. 46) and is a primary source of differences among L2 learners (see reviews by Ganschow & Sparks, 2001; Sparks, 1995). Additional variance in predicting L2 spelling might also be explained by the orthographic distance and orthographic depth between the L1 and the L2. Ellis (2002) has speculated that differences in orthographic transparency are important for the rate of acquisition in another language. Likewise, the orthographic depth of the L2 and the orthographic distance be-



tween the L1 and L2 may affect the rate at which L2 learners acquire spelling skills.

In their study with young ESL learners, Wade-Woolley and Siegel (1997) found that L2 spelling performance was unrelated to the learners' L1 skill and that reading skill was a more significant influence on the ESL learners' spelling accuracy than their L1 skill. In contrast, the results of the present research with older L2 learners showed that L1–L2 spelling skills were strongly correlated (.69). The results suggest that L1–L2 spelling skills are likely to be strongly related even in older L2 learners who had learned to spell (and read) their L1 several years earlier, a finding that supports long-term cross-linguistic transfer of phonological processing skills from L1 to the L2.

### *L2 Reading Comprehension*

The regression analyses demonstrated that among the L1 variables, the measure of L1 reading comprehension (Formal Reading Inventory) was the best predictor of L2 reading comprehension. The finding that L1 reading comprehension was predictive of L2 comprehension is consistent with findings by VanGelderen et al. (2004), who observed that L1 reading comprehension in the 8th grade made large contributions to L2 reading comprehension skills in the 10th grade. The present results extend their research by showing that L1 reading comprehension in elementary school from the 1st through 5th grade was a significant predictor of L2 reading comprehension skill several years later in 10th grade.

The findings show that L1 measures, such as reading comprehension, word decoding, receptive vocabulary, and listening comprehension, explained a small part of the variance (25%) in L2 reading comprehension. Thus, much of the variance in L2 reading comprehension must be explained by other factors. In other studies, researchers have found that L2 word decoding skills play an important role in L2 reading comprehension skills (e.g., see Fukink et al., 2005; Koda, 1998). In this investigation, analyses were also conducted to determine whether first- and second-year L2 decoding skills would be predictive of L2 reading comprehension. The results revealed that when first-year L2 word decoding replaced the Woodcock Basic Skills Cluster as a predictor variable, L1 reading comprehension (Formal Reading Inventory) was still the best predictor of L2 reading comprehension. However, when second-year L2 word decoding replaced first-year L2 word decoding in the analysis, L2 word decoding was the best predictor and explained 39% of the variance ( $\beta = .62$ ) in L2 reading comprehension. This might have been expected because L2 word decoding was administered at the same time as the L2 reading comprehension measure, that is, at the end of the second-year L2 course. Nonetheless, the finding suggests that L2 word decoding explains a substantial part of the variance in L2 reading comprehension and accounts for more of the variance than either L1 reading comprehension skills or L1 word decoding skills. This corroborates studies reviewed by Koda (2005), who concluded that strong decoding skills are a prerequisite for efficient L2 comprehension. The results also support research that has shown that word decoding and other skills related to phonological processing (e.g., spelling) are the primary characteristics that distinguish between good and poor L2 learners (e.g., see Ganschow & Sparks, 2001; Sparks, 1995).

Despite the importance of L2 word decoding skills for L2 reading comprehension, decoding explained less than half of the

variance in L2 reading comprehension. Moreover, L1 language proficiency variables (e.g., L1 receptive vocabulary, L1 listening comprehension) did not explain significant additional variance in L2 reading comprehension. L2 researchers contend that oral proficiency in the L2 becomes more important for L2 reading comprehension as students become more proficient decoders and begin to read more difficult L2 text (Cummins, 1986; Nation, 2001). VanGelderen et al. (2004) found that L2 vocabulary knowledge significantly predicted L2 reading comprehension. In another study, Proctor et al. (2005) cited evidence that showed that the most important indicator of L2 oral proficiency may be L2 vocabulary knowledge, which is important for comprehension of both spoken and written language. In their study with fourth-grade ESL learners, they found that although L2 word decoding skills are important for L2 reading comprehension, word decoding played a less predictive role in L2 reading comprehension than did oral language proficiency and that L2 vocabulary knowledge played a crucial role in improving L2 reading comprehension. Other researchers have verified that the role of L2 vocabulary knowledge becomes more influential among adolescent and adult L2 learners (e.g., see Alderson, 1984; Laufer, 1997).

The aforementioned findings about L2 word decoding and L2 proficiency, especially L2 vocabulary knowledge, suggest that some L2 learners in the United States who do not begin to study an L2 until high school may experience difficulties developing their L2 reading comprehension skills. In the course of reading development in the L1, students read more difficult text as their decoding skills improve. As they read, they enhance their vocabulary and are exposed to increasingly difficult grammatical knowledge and decontextualized language, which also improves comprehension. In high school L2 courses, however, reading is both a mechanism by which instruction is delivered and a target goal of the L2 course (Proctor et al., 2005). Indeed, much of the instruction in high school L2 classrooms is conducted in the students' L1 because students are not yet competent in either the written or oral aspects of the L2. Coady (1997, p. 229) has identified what he called the beginner's paradox, which asks the question, "How can they learn enough words to learn vocabulary through extensive reading when they do not know enough words to read well?" L2 educators have rarely focused on teaching and assessing L2 word decoding skills as a prerequisite for reading L2 text (Koda, 2005); however, if students cannot decode L2 words well for reading, then it is unlikely that they will read very much or that their L2 classroom teachers will require them to read increasingly difficult L2 text. Proctor et al. (2005) cited evidence that indicates that adult L2 reading is similar to L1 reading comprehension processes as L2 word and grammar knowledge improve. However, it is unlikely that high school students who attempt to add an L2 in high school courses primarily as one of several academic subjects will become proficient in L2 reading comprehension if they do not acquire adequate L2 word decoding skills and do not read increasingly difficult text as an integral and ongoing part of their L2 instruction.

### *Limitations and Implications*

The present study is unique in several ways. In particular, the participants were followed over 10 years so that the possibility of long-term cross-linguistic transfer of L1 reading skills to L2 reading could be examined. Also, the participants learning an L2 were



high school students who were already proficient in the oral and written aspects of their L1 and generally were not studying the L2 to become bilingual but instead to fulfill an academic requirement. However, the study also has limitations that restrict generalizing the findings. For example, the small sample size limits the power of the statistical analyses. In addition, the use of nonvalidated measures for assessing L2 word decoding, spelling, and reading comprehension was necessary because standardized instruments for measuring these skills in an L2 are unavailable. Attempts were made to construct L2 assessments that closely resembled the L1 measures of this research; for example, on both the L1 and L2 reading comprehension tests, the student read the text silently and answered multiple-choice comprehension questions. Although the internal consistency of the L2 measures was checked, their construct or criterion-related validity was not determined. Thus, it is possible that higher scores on any one of the L2 measures, with the other two scores held constant, would not necessarily result in higher L2 proficiency in L2 reading. Also, Koda (2005) has proposed that for L2 reading comprehension, one type of assessment task, for example, multiple-choice items, cannot measure all of the skills necessary for successful comprehension. In addition, the participants studied three different L2s, each of which has a different orthography, and their scores were combined in the analyses. Though the three L2s were alphabetic orthographies, Koda has cautioned that "transferred skills [from L1 to L2] continuously undergo adjustments to accommodate L2 orthographic peculiarities" (Koda, 2005, p. 246). Therefore, combining the three different L2 groups may have served to blur unique distinctions between L1-L2 decoding, spelling, and reading comprehension in the three L2s.

Nonetheless, there are several ways in which the study contributes to both the L1 and L2 reading literature. First, mastery of L1 decoding skills early in the primary school years may be beneficial for students who attempt to learn an L2 in high school. Second, direct and explicit teaching of the phonology and orthography of the L2 to students may be beneficial in learning to read and spell the L2, especially if the L1 was mastered several years earlier. Likewise, direct and explicit teaching of the grammar of the L2 may be beneficial for comprehending an L2, especially if there are distinct differences between syntactic processing requirements of the L1 and L2. Third, after learning to decode the L2, new L2 learners may benefit from reading as much L2 text as possible because they would be more likely to increase their fluency, acquire the vocabulary and grammar of the L2, and learn the background knowledge they will need to comprehend the L2. Finally, L2 educators and researchers can collaborate to study further the importance of decoding skills for mastery of L2 reading.

## References

- Aaron, P. G. (1989). Can reading disabilities be diagnosed without using intelligence tests? *Journal of Learning Disabilities*, 24, 178-191.
- Akamatsu, N. (1999). The effects of first language orthographic features on word recognition processing in English. *Reading and Writing: An Interdisciplinary Journal*, 11, 381-403.
- Alderson, J. (1984). Reading in a foreign language: A reading problem or a language problem? In J. Alderson & A. Urquhart (Eds.), *Reading in a foreign language* (pp. 1-24). London: Longman.
- American Council on the Teaching of Foreign Languages. (1986). *ACTFL proficiency guidelines*. Yonkers, NY: Author.
- American Council on the Teaching of Foreign Languages. (1989). *American Council on the Teaching of Foreign Languages proficiency guidelines*. Hastings-on-Hudson, NY: Author.
- August, D., Calderon, M., & Carlo, M. (2001). *Transfer of skills from Spanish to English: A study of young learners*. Washington, DC: Center for Applied Linguistics.
- Bernhardt, E. (1991). *Reading development in a second language*. Norwood, NJ: Ablex.
- Bialystock, E. (2001). *Bilingualism in development*. Cambridge, England: Cambridge University Press.
- Bradley, L., & Bryant, P. (1983, February 3). Categorizing sounds and learning to read—A causal connection. *Nature*, 301, 419-421.
- Brown, T., & Haynes, M. (1985). Literacy background and reading development in a second language. In T. Carr (Ed.), *The development of reading skills* (pp. 19-34). San Francisco: Jossey-Bass.
- Bruck, M., & Treiman, R. (1990). Phonological awareness and spelling in normal children and dyslexics: The case of initial consonant clusters. *Journal of Experimental Child Psychology*, 50, 156-178.
- Carroll, J., & Sapon, S. (1959). *Modern Language Aptitude Test (MLAT): Manual*. San Antonio, TX: Psychological Corporation.
- Chamot, A., & El-Dinary, P. (1999). Children's learning strategies in language immersion classes. *Modern Language Journal*, 83, 319-338.
- Cisero, C., & Royer, J. (1995). The development of cross-language transfer of phonological awareness. *Contemporary Educational Psychology*, 20, 275-303.
- Clarke, M. (1980). The short circuit hypothesis of ESL reading: Or when language competence interferes with reading performance. *Modern Language Journal*, 64, 203-209.
- Coady, J. (1997). L2 vocabulary acquisition through extensive reading. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 225-237). Cambridge, England: Cambridge University Press.
- Comeau, L., Cormier, P., Grandmaison, E., & Lacroix, D. (1999). A longitudinal study of phonological processing skills in children learning to read a second language. *Journal of Educational Psychology*, 91, 29-43.
- Cummins, J. (1979). Linguistic interdependence and educational development of bilingual children. *Review of Educational Research*, 49, 222-251.
- Cummins, J. (1984). Implications of bilingual proficiency for the education of minority language students. In P. Allen, M. Swain, & C. Brumfit (Eds.), *Language issues and education policies: Exploring Canada's multilingual resources* (pp. 21-34). Oxford, England: Pergamon.
- Cummins, J. (1986). Empowering minority students: A framework for intervention. *Harvard Educational Review*, 56, 18-36.
- Cummins, J., & Mulchay, R. (1978). Orientation to language in Ukrainian-English bilingual children. *Child Development*, 49, 1239-1242.
- Cunningham, A., & Stanovich, K. (1997). Early reading acquisition and its relation to reading ability 10 years later. *Developmental Psychology*, 33, 934-945.
- Cutting, L., & Scarborough, J. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other language skills can depend on how comprehension is measured. *Scientific Studies of Reading*, 10, 277-299.
- Cziko, G. A. (1980). Language competence and reading strategies: A comparison of first- and second-language oral reading errors. *Language Learning*, 30, 101-114.
- Dufva, M., & Voeten, M. (1999). Native language literacy and phonological memory as prerequisites for learning English as a foreign language. *Applied Psycholinguistics*, 20, 329-348.
- Dunn, L., & Dunn, L. (1981). *Peabody Picture Vocabulary Test—Revised*. Circle Pines, MN: American Guidance.



- Durgunoglu, A. (2002). Cross-language transfer in literacy development and implications for language learners. *Annals of Dyslexia*, 52, 189–204.
- Ehri, L. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading*, 9, 167–188.
- Ellis, N. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, 24, 143–188.
- Favreau, M., & Segalowitz, N. (1982). Second language reading in fluent bilinguals. *Applied Psycholinguistics*, 3, 329–341.
- Favreau, M., & Segalowitz, N. (1983). Automatic and controlled processes in the first- and second-language reading of fluent bilinguals. *Memory & Cognition*, 11, 565–574.
- Frith, U. (1980). *Cognitive processes in spelling*. New York: Academic Press.
- Fukink, R., Hulstijn, J., & Sims, A. (2005). Does training in second-language word recognition skills affect reading comprehension? An experimental study. *Modern Language Journal*, 89, 54–75.
- Ganschow, L., & Sparks, R. (2001). Learning difficulties and foreign language learning: A review of research and instruction. *Language Teaching*, 34, 79–98.
- Geva, E., & Verhoeven, L. (2000). Introduction: The development of second language reading in primary children—Research issues and trends. *Scientific Studies of Reading*, 4, 261–266.
- Geva, E., Wade-Woolley, L., & Shany, M. (1997). Development of reading efficiency in first and second language. *Scientific Studies of Reading*, 1, 119–144.
- Geva, E., Yaghou-Zadeh, Z., & Schuster, B. (2000). Understanding individual differences in word recognition skills of ESL children. *Annals of Dyslexia*, 50, 121–154.
- Gholamain, M., & Geva, E. (1999). Orthographic and cognitive factors in the concurrent development of basic reading skills in English and Persian. *Language Learning*, 49, 183–217.
- Haynes, M., & Carr, T. (1990). Writing system background and second language reading: A component skills analysis of English reading by native-speaking readers of Chinese. In T. Carr & B. Levy (Eds.), *Reading and its development: Component skills approaches* (pp. 375–421). San Diego, CA: Academic Press.
- Hirai, A. (1994). The relationship between listening and reading rates of Japanese EFL learners. *Modern Language Journal*, 83, 367–384.
- Holm, A., & Dodd, B. (1996). The effect of first written language on the acquisition of English literacy. *Cognition*, 59, 119–147.
- Horiba, Y. (1990). Narrative comprehension processes: A study of native and non-native readers of Japanese. *Modern Language Journal*, 74, 188–202.
- Hulstijn, J., & Bossers, B. (1992). Individual differences in L2 proficiency as a function of L1 proficiency. *European Journal of Cognitive Psychology*, 4, 341–353.
- Humes-Bartlo, M. (1989). Variation in children's ability to learn second languages. In K. Hyltenstam & L. Obler (Eds.), *Bilingualism across the life span* (pp. 41–54). Cambridge, England: Cambridge University Press.
- Kahn-Horwitz, J., Shimron, J., & Sparks, R. (2005). Predicting foreign language reading achievement in elementary school students. *Reading and Writing: An Interdisciplinary Journal*, 18, 527–558.
- Kahn-Horwitz, J., Shimron, J., & Sparks, R. (2006). Weak and strong novice readers of English as a foreign language: Effects of first language and socioeconomic status. *Annals of Dyslexia*, 56, 161–185.
- Katz, L., & Frost, R. (1992). Reading in different orthographies: The orthographic depth hypothesis. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 67–84). Amsterdam: Elsevier.
- Koda, K. (1990). The use of L1 reading strategies in L2 reading. *Studies in Second Language Acquisition*, 12, 393–410.
- Koda, K. (1992). The effects of lower-level processing skills on FL reading performance: Implications for instruction. *Modern Language Journal*, 76, 502–512.
- Koda, K. (1998). L2 word recognition research: A critical review. *Modern Language Journal*, 80, 450–460.
- Koda, K. (1999). Development of L2 intraword structural sensitivity and decoding skills. *Modern Language Journal*, 83, 51–64.
- Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge, England: Cambridge University Press.
- Larsen, S., & Hammill, D. (1986). *Test of Written Spelling—2 (TWS-2)*. Austin, TX: PRO-ED.
- Laufer, R. (1997). The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 20–34). Cambridge, England: Cambridge University Press.
- Legarretta, D. (1979). The effects of program models on language acquisition of Spanish speaking children. *TESOL Quarterly*, 13, 521–534.
- Lindamood, L., & Lindamood, P. (1973). *Lindamood Auditory Conceptualization Test*. Allen, TX: PRO-ED.
- Lindsey, K., Manis, F., & Bailey, C. (2003). Prediction of first-grade reading in Spanish-speaking English-language learners. *Journal of Educational Psychology*, 95, 482–494.
- Lundberg, I., Olofsson, A., & Wall, S. (1980). Reading and spelling skills in the first years predicted from phonemic awareness skills in kindergarten. *Scandinavian Journal of Psychology*, 21, 159–173.
- MacDonald, G., & Cornwall, A. (1995). The relationship between phonological awareness and reading and spelling achievement eleven years later. *Journal of Learning Disabilities*, 28, 523–527.
- Mendenhall, W., & Sincich, T. (1989). *A second course in business statistics* (3rd ed.). San Francisco: Dellen.
- Meschyan, G., & Hernandez, A. (2002). Is native-language decoding skill related to second-language learning? *Journal of Educational Psychology*, 94, 14–22.
- Moats, L. (2005). How spelling supports reading. *American Educator*, Winter, 12–22, 42–43.
- Muljani, D., Koda, K., & Moates, D. (1998). The development of word recognition in a second language. *Applied Psycholinguistics*, 19, 99–113.
- Nassaji, H., & Geva, E. (1999). The contribution of phonological and orthographic processing skills to adult ESL reading: Evidence from native speakers of Farsi. *Applied Psycholinguistics*, 20, 241–267.
- Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge, England: Cambridge University Press.
- Olshtain, E., Shohamy, E., Kemp, J., & Chatow, R. (1990). Factors predicting success in EFL among culturally different learners. *Language Learning*, 40, 23–44.
- Proctor, C., Carlo, M., August, D., & Snow, C. (2005). Native Spanish-speaking children reading in English: Toward a model of comprehension. *Journal of Educational Psychology*, 97, 246–256.
- Rayner, K., Foorman, B., Perfetti, C., Pesetsky, D., & Seidenberg, M. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest*, 2, 31–74.
- Ryan, A., & Meara, P. (1991). The case of invisible vowels: Arabic speakers reading English words. *Reading in a Foreign Language*, 7, 531–540.
- Scott, Foresman and Company. (1987). *Test of Reading Readiness*. Glenview, IL: Author.
- Segalowitz, N. S. (1986). Skilled reading in the second language. In J. Vaid (Ed.), *Language processing in bilinguals: Psycholinguistic and neurological perspectives* (pp. 3–19). Hillsdale, NJ: Erlbaum.
- Segalowitz, N. S., Poulsen, C., & Komoda, M. (1991). Lower level components or reading skill in higher level bilinguals: Implications for reading instruction. *AILA Review*, 8, 15–30.
- Snow, C., Burns, M., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.
- Sparks, R. (1995). Examining the linguistic coding differences hypothesis

- to explain individual differences in foreign language learning. *Annals of Dyslexia*, 45, 187–214.
- Sparks, R., Artzer, M., Patton, J., Ganschow, L., Miller, K., Hordubay, D., & Walsh, G. (1998). Benefits of multisensory language instruction in Spanish for at-risk learners: A comparison study of high school Spanish students. *Annals of Dyslexia*, 48, 239–270.
- Sparks, R., & Ganschow, L. (1993). Searching for the cognitive locus of foreign language learning problems: Linking first and second language learning. *Modern Language Journal*, 77, 289–302.
- Sparks, R., Ganschow, L., Artzer, M., Siebenhar, D., & Plageman, M. (1998). Differences in native language skills, foreign language aptitude, and foreign language grades among high, average, and low proficiency learners: Two studies. *Language Testing*, 15, 181–216.
- Sparks, R., Humbach, N., & Javorsky, J. (in press). Comparing high and low achievement, LD, and ADHD foreign language learners: Individual and longitudinal differences. *Learning and Individual Differences*.
- Sparks, R., Patton, J., Ganschow, L., Humbach, N., & Javorsky, J. (2006). Native language predictors of foreign language proficiency and foreign language aptitude. *Annals of Dyslexia*, 56, 129–160.
- Stuart, M., & Masterson, J. (1992). Patterns of reading and spelling in 10-year-olds related to prereading phonological abilities. *Journal of Experimental Child Psychology*, 54, 168–187.
- Treiman, R. (1993). *Beginning to spell: A study of first-grade children*. New York: Oxford University Press.
- VanGelderens, A., Schoonen, R., deGlopper, K., Hulstijn, J., Simis, A., Snellings, P., & Stevenson, M. (2004). Linguistic knowledge, processing speed, and metacognitive knowledge in first- and second-language reading comprehension: A componential analysis. *Journal of Educational Psychology*, 96, 19–30.
- Verhoeven, L. (2000). Components in early second language reading and spelling. *Scientific Studies of Reading*, 4, 313–330.
- Wade-Woolley, L., & Siegel, L. (1997). The spelling performance of ESL children and native speakers of English as a function of reading skills. *Reading and Writing: An Interdisciplinary Journal*, 9, 387–406.
- Wagner, R., & Torgesen, J. (1987). The nature of phonological processing and its causal role in reading acquisition. *Psychological Bulletin*, 101, 192–212.
- Wiederholt, L. (1986). *Formal Reading Inventory*. Austin, TX: Pro-Ed.
- Wolf, M. (Ed.). (2001). *Dyslexia, fluency, and the brain*. Timonium, MD: York Press.
- Woodcock, R. (1987). *Woodcock Reading Mastery Test—Revised*. Circle Pines, MN: American Guidance.

## Appendix A

### Lists of Real Words and Pseudowords for Spanish, French, and German

Spanish		French		German	
Real words	Pseudowords	Real words	Pseudowords	Real words	Pseudowords
anoche	loche	adorable	elars	frei	walstatt
enero	regua	midi	trousante	schier	angesammen
isla	traceo	robe	etez	Nachbar	biedel
orilla	placeta	chaise	pentants	bitte	leiner
usted	sucrete	difficile	tonseur	Bücher	nochtbar
mesa	popeta	croix	pateau	Einwand	auwinnern
señora	porrosca	ouest	gaissons	Todesurteile	sotürlich
jefa	asurge	soeur	pretons	Pflicht	zeiben
entretenimiento	hastilla	magnifique	els	jawohl	speulen
salón	movadiza	aéroport	lace	deshalb	möffel
inventado	vestuto	soleil	desentrez	ausgezeichnet	besuchgekommen
mural	cantón	joël	tadelle	Fussboden	kössen
la amada	calahoria	longtemps	ponte	leute	weschen
agencia	meradario	pluie	exploter	Flugzeug	hilgescher
filólogo	zebajo	médicament	d'assoint	besonders	hols
repentinamente	pantaora	rançon	soulangerie	zogen	wangst
antena	cebaduría	hier	trençons	Geschlecht	vertaunen
corriente	grallanado	décidé	boignon	Räuber	mautern
alto	llenosidad	nationalité	ramposé	rauchen	leisling
preocupación	yagüe	réveillon	touvertes	unbezähmbar	dömpferbehn

(Appendixes continue)



## Appendix B

## List of Spelling Words for Spanish, French, and German

Spanish	French	German
bien	faire	Zeit
arpa	comme	Handschuh
dulce	maison	natürlich
gafas	cher	Lederhosen
Litro	vraiment	Kuchen
maíz	voudrais	Bücher
pago	étoile	verbinden
después	pendant	hoffentlich
secreto	visiter	Wechsel
paisaje	poivre	Bahnhof
placita	déjeuner	Bienenstich
tranquilo	choisir	pferde
mirando	pourrais	gefallen
debajo	adorable	schmeckt
horarios	printemps	Frucht
tagarote	chanteuse	gewöhnlich
periódico	raconteur	besuchen
felicidad	magnifique	Schule
zanahoria	chaise	Obst
abecedario	ouvrir	Mittagessen

Received May 23, 2006  
Revision received July 9, 2007  
Accepted August 11, 2007 ■

# The Mnemonic Value of Orthography for Vocabulary Learning

Julie Rosenthal and Linnea C. Ehri  
Graduate Center of the City University of New York

In 2 experiments, the authors examined whether spellings improve students' memory for pronunciations and meanings of new vocabulary words. Lower socioeconomic status minority 2nd graders ( $M = 7$  years 7 months;  $n = 20$ ) and 5th graders ( $M = 10$  years 11 months;  $n = 32$ ) were taught 2 sets of unfamiliar nouns and their meanings over several learning trials. The words were defined, depicted, and embedded in sentences. During study periods, students were shown written forms of 1 set but not the other set. Spellings were not present during word recall. Results of analyses of variance showed that spellings enhanced memory for pronunciations and meanings compared to no spellings ( $ps < .01$ ). Better readers and spellers increasingly outdistanced poorer readers and spellers in remembering pronunciations over trials when spellings accompanied learning ( $p < .05$ ), suggesting a Matthew effect. An explanation is that spellings activated graphophonemic connections to better secure pronunciations and meanings in memory. Results indicate that orthographic knowledge benefited vocabulary learning and diminished dependence on phonological memory. Instructional implications are that teachers should include written words as part of vocabulary instruction and that students should pronounce spellings as well as determine meanings when they encounter new vocabulary words.

**Keywords:** vocabulary learning, orthographic memory, phonological working memory, spelling, vocabulary instruction

Vocabulary knowledge is central to successful reading comprehension and academic achievement (Beck & McKeown, 1999; National Reading Panel, 2000). Children's vocabularies grow rapidly during the school years. However, the variation in vocabulary growth among children is great. One factor contributing to this variation is the well-documented disparity between vocabularies of advantaged and disadvantaged populations (Chall, Jacobs, & Baldwin, 1990; Hart & Risley, 1995; Snow, Burns, & Griffin, 1998). Findings reveal that children who come to school with limited vocabularies never catch up to their higher quartile peers (Biemiller, 2001). In fact, students who are below average in their vocabularies and reading ability experience a reading slump in Grade 4, when subject matter texts become increasingly abstract and complex. This slump worsens in seventh grade and beyond, when knowledge of word meanings becomes crucial to academic success (Chall et al., 1990). Thus, there is a clear need for better methods of building students' vocabularies in school. The purpose of the present study was to examine the contribution of ortho-

graphic knowledge—that is, exposure to the spellings of words—in facilitating students' vocabulary growth.

Vocabulary knowledge consists of the stock of words known by a person. Word knowledge involves multiple identities and dimensions (Nagy & Scott, 2000): a phonological identity including constituent phonemes and how the word is pronounced; one or more syntactic identities specifying the word's form class and function in sentences; one or more semantic identities, or meanings of the word in various contexts; and an orthographic identity, or the word's spelling. This knowledge may also include morphological information distinguishing roots and affixes or compound parts of multimorphemic words. It may include other words that typically co-occur with the word or are associated with it. Its meaning may include not only referential elements that are visual and verbal but also affective elements. Word knowledge may range from partial to complete, or it may involve misconceptions. It may support only recognizing word meanings, or it may enable productive use of the word. A vocabulary learning task may include a word's formal definition, its use in various sentence contexts, or a picture of its referent. It may require learning a new concept or learning a new label for a known concept. Vocabulary learning may result from explicit instruction by a teacher, or it may emerge from incidental encounters with unknown words. One may learn words and their meanings by hearing them spoken or by reading them. However, simply learning to decode or to spell words or to read words by sight when the meanings of those words are already familiar or when meanings are lacking (i.e., nonwords) is not considered part of vocabulary learning.

Although there are multiple dimensions to vocabulary knowledge, its essence involves remembering the pronunciations of words and their meanings. According to Gathercole (2006) and Gathercole and Baddeley (1993), being able to repeat a new

---

Julie Rosenthal and Linnea C. Ehri, Program in Educational Psychology, Graduate Center of the City University of New York.

The study was funded in part by the Jeanne S. Chall Research Fellowship from the International Reading Association and by the Mario Capelloni Dissertation Fellowship from the Graduate Center of the City University of New York. We thank Hollis Scarborough and David Rindskopf for their helpful comments and suggestions. We are grateful to students and staff at the public elementary school in New Jersey where this study was conducted.

Correspondence concerning this article should be addressed to Julie Rosenthal, who is now at the Department of Elementary and Early Childhood Education, William Paterson University of New Jersey, 300 Pompton Road, Wayne, NJ 07470. E-mail: rosenthalj@wpunj.edu



word's pronunciation accurately depends on having good phonological short-term working memory and is critical for explaining how people, especially young children, are able to remember spoken words. When words are repeated, they become represented temporarily in a phonological loop in working memory until they can be transferred to a more permanent lexical-semantic memory system. Phonological working memory is assessed with a nonword repetition task. In one study, Gathercole and Baddeley (1989) showed that phonological working memory capacity predicted the ease with which children between the ages of 4 and 6 years learned novel phonological blends as names for known objects, but it did not predict learning when the names were familiar. However, among 8-year-olds, nonword repetition was much less predictive of vocabulary scores. Other studies have confirmed that individual differences in phonological working memory underlie vocabulary acquisition, particularly in young children (Gathercole, 2006). The purpose of the present study was to examine whether exposure to the spellings of words might impact the formation of phonological representations of new vocabulary words in memory.

Research on vocabulary acquisition has been extensive (see Beck & McKeown, 1991, 1999; Biemiller, 2006; Nagy & Scott, 2000). One of the puzzles has been to explain vocabulary growth. Students acquire knowledge of many more words than they are taught explicitly by teachers or look up in dictionaries. The consensus is that much of this knowledge is acquired incidentally through exposure to words in speech or print. Many vocabulary studies conducted with students beyond the early years, after they have learned to read, have examined word learning through print exposure (Beck & McKeown, 1991; Swanborn & deGlopper, 1999). However, few, if any, studies to date have examined experimentally whether seeing the spellings of words contributes to vocabulary learning. The prevailing view appears to be that when new vocabulary words are encountered in print, they are transformed into pronunciations, which are then stored in memory along with word meanings. The written form is not viewed as making any special contribution to memory beyond its spoken form. Gathercole and Baddeley (1993) did not include orthographic representations in their model.

Examination of recommendations by vocabulary instruction experts reveals little, if any, explicit incorporation of print into instruction. Blachowicz and Fisher (2004) recommended that teachers expand students' vocabularies by reading text aloud to their classes, thereby providing access to words too difficult for students to read independently. They suggested that teachers stop at difficult words, provide definitions, reread the text to "cement" new word meanings, and have students "act out" the meanings of new words. However, they did not suggest displaying the spellings of new words. Biemiller (2004) also recommended that teachers read texts aloud to students in kindergarten through Grade 2 to build their vocabularies. He recommended using books with more sophisticated vocabulary and explaining harder words during rereadings, but nowhere did he mention exposing students to spellings of words. Trelease (2006), in his very popular *The Read Aloud Handbook*, stated that the single most effective way to build students' vocabularies is by reading aloud to them, but he, too, never suggested showing word spellings. Beck, McKeown, and Kucan (2002; Beck & McKeown, 2007) recommended multiple steps for teaching new words from trade books after teachers have read them aloud to students: (a) review the more difficult words

and where they occurred in the story and explain their meanings, (b) have children repeat the words aloud to build their phonological representations and orally present the words in sentences, (c) have children construct additional sentences, and (d) then have them say the words aloud again. At the very end, the spellings of the words are finally shown when the words are added to a word wall in the classroom.

Whereas orthography has been overlooked in studies of vocabulary acquisition, the meanings of words have been overlooked in research on the acquisition of decoding skill. In this work, the primary interest has been on the formation of connections between orthography and phonology that enable children to decode and spell written words rather than on the formation of semantic connections. In fact, decoding researchers regard the ability to read nonwords, which have no meaning, as the best way to assess readers' skill in reading unfamiliar (i.e., new vocabulary) words (Ehri, 1991; Gough & Tunmer, 1986; Rack, Snowling, & Olson, 1992; Share, 1995). As a result, the vocabulary and decoding research domains overlap very little. One purpose of the present study was to bring the two closer together by examining whether and how orthography contributes to the acquisition of words and their meanings.

In the present study, vocabulary learning consisted of explicitly teaching children the spoken forms and meanings of several monomorphemic nouns. The words were pronounced, defined, and embedded in sentences, and drawings of referents were displayed on flash cards to teach pronunciations, syntactic roles, meanings, relationships to other words, and links to world knowledge. In addition, spellings of the words appeared on the cards in one condition, while spellings were not seen in the other condition. We assessed vocabulary learning by evaluating whether children could recall pronunciations of the words when prompted with their meanings and could recall their meanings when they heard the words spoken.

There are various reasons to expect that exposure to the written forms of unfamiliar words might help students learn and remember them better. According to Ehri's (1992, 1999, 2005) connectionist theory, knowledge of the orthographic writing system provides a powerful mnemonic device enabling readers to secure the spellings of specific words to their pronunciations in memory. For connections to be formed, readers must be able to distinguish phonemes in spoken words, know how graphemes symbolize those phonemes systematically in the writing system, and compute these grapho-phonemic connections when they see spellings and speak or hear pronunciations of the words. As a result of this processing, spellings of the words become bonded to their pronunciations and meanings, and the amalgam is stored in memory. Having words stored in memory precludes the need to apply sounding out, analogy, or guessing strategies to read the words. On seeing familiar words, readers can recognize them from memory (by sight). Others have proposed similar theories (Rack, Hulme, Snowling, & Wightman, 1994; Perfetti, 1992; Share, 1995, 1999).

Ehri and Wilce (1979) showed that spellings function as a mnemonic device enabling readers to retain pronunciations of nonwords in memory. An everyday example of this function is when one hears an unfamiliar personal name, examines how it is spelled, and remembers it better as a result. In Ehri and Wilce's study, first and second graders were given several trials to learn four spoken nonsense syllables (e.g., *jad*, *sim*, *des*, *fug*). The



conditions of learning varied across groups. The experimental group was shown written forms of the words during study periods but not during recall trials. Several control groups were included. The students in these groups rehearsed the spoken forms of nonwords extra times in various ways but did not see their spellings. Across all comparisons, students recalled the nonwords significantly better when they had been exposed to spellings during study periods than when they had not. The effect of spelling aids on memory was stronger for students with larger lexicons of printed words, which indicates that better orthographic knowledge strengthened the benefit of spellings. Results were interpreted to support the idea that spellings facilitate word learning because they activate the formation of orthographic images symbolizing and securing the words' pronunciations in memory. However, because nonwords were used, findings provided no information about the learning of new words that included meanings.

The purpose of the present study was to adapt the word-learning procedures used by Ehri and Wilce (1979) to study vocabulary learning. The question of interest was whether elementary students would better learn and remember the pronunciations and meanings of new words when they were exposed to spellings of the words than when they practiced only spoken forms of the words. A paired associate learning task was adopted to study these processes in a controlled way. In a repeated measures, counterbalanced design with random assignment, second and fifth graders were taught two sets of low-frequency nouns and their meanings. For one set, they saw spellings of the words during study periods. For the other set, they did not see the spellings. Instead, they received extra practice hearing and pronouncing the words. Use of this control for exposure to spellings was based on the prevailing view of print in vocabulary learning, that spellings of new words are transformed into pronunciations, which are the forms associated with meanings in memory. It is important to note that when recall of pronunciations was tested, written words were never present, so children had to retrieve them from memory.

This design was used to test several hypotheses. On the basis of grapheme-phoneme connectionist theories, we reasoned that spelling exposure should facilitate learning pronunciations, because spellings better secure pronunciations in memory. We also expected that word meanings would be learned more quickly if spellings secured pronunciations in memory earlier during the learning trials. Regarding grade-level differences, we reasoned that the benefit of orthography might be diminished in fifth graders compared to second graders because older students possess larger vocabularies, more extensive memory for the phonological forms of words, and greater decoding proficiency.

Extensive research on vocabulary learning has clarified factors that influence the ease of learning word meanings. Carey (1978) observed "fast mapping" in preschool children, who could very quickly learn the referents for new words but took longer to acquire a more complete understanding of their meanings. Vocabulary learning is faster when concrete images of word meanings can be represented in memory and when words are defined verbally and experienced in multiple sentence contexts, so that their meanings are connected to other word meanings, becoming embedded in a speaker's web of linguistic and world knowledge, as portrayed by dual coding theory (Sadoski, 2005). The present study included pictures, definitions, and sentences as part of vocabulary learning to support the acquisition of word meanings.

Many vocabulary studies have examined growth in receptive knowledge, most often measured by multiple-choice tests (Senechal, 1997). However, language users need to know new vocabulary words well enough to incorporate them into their speech. The requirements for doing this are more stringent than the requirements for simply recognizing words because the latter process may be aided by morphological information, verbal context, or guessing (Baumann & Kameenui, 1991). The present study was intended to examine whether seeing the spellings of new words strengthens growth in expressive vocabulary.

## EXPERIMENT 1: VOCABULARY LEARNING WITH SECOND GRADERS

### Method

#### *Participants*

In a mid-size city, 20 second graders were selected from a school where 58% of the students qualified for free or reduced-price lunch. Participants were those whose parents gave written consent, who were not receiving English language services, and who exhibited no cognitive or sensory impairments according to their teachers. There were 11 girls and 9 boys, of whom 15 were African American and 5 were Latin American, with a mean chronological age of 7 years 7 months, tested in the fall.

#### *Materials and Procedures*

##### *Literacy Assessment*

Several tests were administered. The purpose was to assess students' orthographic knowledge and vocabulary level.

*Reading words.* Graded lists of words on the Word Identification subtest of the Woodcock Reading Mastery Test—Revised (WRMT-R; Woodcock, 1987) were given to assess students' word-reading level. According to the manual, the reliability is .98.

*Reading and spelling nonwords.* Nonsense words were constructed from letters appearing in the new vocabulary words to be taught. Students read 15 consonant-vowel-consonant (CVC) nonwords and spelled 8 CVC nonwords containing short vowels. Examples are *pag*, *kem*, *hin*, *nol*, and *yud*. The split-half reliabilities in our sample were .89 (nonword reading) and .28 (nonword spelling). One reason for the lower value in the latter case is that there were only 8 nonwords, and at least 3 of these were misperceived as real words by some students (e.g., *hin* spelled *hen*).

*Vocabulary.* The Peabody Picture Vocabulary Test (PPVT-III; Dunn & Dunn, 1997) was administered to assess students' receptive vocabulary. According to the manual, reliabilities range from .91 to .94.

##### *Vocabulary Learning*

Two sets of six concrete, low-frequency CVC nouns served as the target vocabulary words, one set to be taught with spellings, the other set to be taught without spellings in a repeated measures design. We consulted several "rare words" Web sites to obtain the words: <http://hometown.aol.com/rlongman1/wordAO.html>, <http://www.islandnet.com/~egbird/dict/dict.htm>, and <http://phrontistery.info/>.



Teachers of the students were shown the words and agreed that their students were unlikely to know them. All but one word (*lad*) was rated as low frequency (i.e., occurring less than 11 times per million words) by Thorndike and Lorge (1972). Although *lad* was more frequent, Harris and Jacobson (1982) rated it as a fourth-grade-level word.

Words in each set began with a different consonant and included a different short vowel, except for short *a*, which appeared twice in each set. A core definition was constructed that included a close synonym for each word. These appear in Table 1. Some of the meanings were simplified. For example, *nib*, meaning the point of a pen, was taught as "a pen." The spelling of one word, *fete*, was regularized to *fet*. Five meaning-clarifying sentences were written for each word—for example,

A keg can hold many kinds of liquids. A keg is big and can hold a lot of water. A keg is usually made out of wood or plastic. A keg is big and round. Some people keep pickles in a keg.

Pictures depicting the meanings of the nouns were drawn on cards. One set of cards showed the written words beneath pictures, while the other set did not. Another set of cards showed only the written words.

At the start of the vocabulary learning tasks, procedures were explained and illustrated with an example. Children were told that they should learn the name and meaning of each picture because later they would be tested. The first trial was a study trial, when the words were introduced. Students were shown each picture, heard the word pronounced and embedded in a defining sentence, and repeated the word and sentence. All subsequent trials prompted their recall of pronunciations and meanings. Two types of prompted recall trials were interleaved during learning: a trial presenting all six pictures and testing memory for spoken words, followed by a trial presenting the six spoken words and testing memory for their definitions. The order of the words was varied across trials.

During pronunciation recall trials, each picture depicting the word was presented, and children were given 5 s to say the word. Feedback followed in which the experimenter continued to display the picture, spoke the word, and embedded it in one or two sentences. On Test Trial 1, the embedded sentence was the defining sentence. On subsequent trials, different meaning-elaborated sentences followed the defining sentence. If children misrecalled an item, they repeated the word and its defining sentence.

During meaning recall trials, each word was spoken, and children had 5 s to recall the definition. Feedback consisted of the experimenter providing the defining sentence. After the first test trial, feedback on each trial included a different meaning-elaborated sentence. Again, if children misrecalled a word's definition, they repeated the word and the defining sentence.

Oral definitions were scored as correct or incorrect. There were no instances of nearly correct responses because all of the words were concrete nouns that were close synonyms of familiar nouns and were included in the defining sentences. Students' responses were correct if they produced the more familiar synonym. For example, the word *cur* was defined as a homeless dog. Children had to produce the close synonym *dog* to be correct. Errors on this task were mostly recalling synonyms of other target words or no response.

Each student learned two sets of words. One set was taught by exposure to spellings of the words. However, no attention was drawn to the written words during the experiment, and the written words were never present when children recalled their spoken forms, so that any influence of spellings arose from students' memory for the spoken words. Exposure to spellings was limited to specific points during the study and test trials. Written words appeared beneath pictures of the nouns during the initial study trial. During test trials, after each attempt to recall the spoken word, the experimenter provided feedback by showing the written words beneath their pictures while saying the words and their sentences. Recall of the meanings of words was prompted by the

Table 1

*Mean Number of Second Graders Who Recalled Spoken Words and Meanings Correctly Per Trial During the First Five Learning Trials for Each Vocabulary Word as a Function of Whether Students Were Exposed to Spellings of the Words in Experiment 1*

Vocabulary word	Meaning	Recall spoken word		Recall meaning	
		Sp.	No sp.	Spell	No sp.
Tot	A small, young child	7.2	5.0	10.0	8.4
Hun	Someone who destroys things	6.6	5.2	9.6	9.6 <sup>a</sup>
Lad	A boy who works with horses	6.0	4.0	9.0	8.0
Sod	Wet, grassy ground	5.2	3.4	8.0	7.8
Pap	Soft, mushy food for babies	5.0	3.2	8.2	8.8 <sup>a</sup>
Cur	A homeless dog	5.0	2.6	9.6	8.4
Yag	Fake jewelry	4.6	2.6	8.4	7.6
Fet	A big, fun party	4.2	2.4	8.8	6.8
Nib	A pen	4.0	1.4	8.8	8.4
Gam	A family of whales	3.8	2.2	9.0	6.0
Keg	A barrel that holds water	3.6	1.8	8.4	8.2
Rig	A big, strong truck	3.2	1.6	8.8	6.4

*Note.* The maximum number of students who recalled a word correctly was 10. Sp. = spelling.

<sup>a</sup> There were two exceptions to the general pattern of superior learning when spellings were seen compared with when spellings were not seen, both involving the recall of word meanings.

experimenter saying the words while showing their spellings for less than 1 s as the word was spoken.

In the spelling-absent condition, no written words accompanied spoken words, sentences, or pictures. Instead, the experimenter pronounced the words additional times, and children repeated the words additional times to equate stimulus exposures across the two conditions. In fact, the total number of oral exposures to the words in the spelling-absent condition was greater than the total number of exposures (written plus oral) in the spelling-present condition. Children heard and repeated each word twice during every feedback trial after attempting to recall its pronunciation, and they heard each word twice rather than once on test trials as they attempted to recall its definition.

Children received a minimum of six and maximum of nine trials to learn pronunciations and definitions in each condition. If they reached a criterion of three perfect consecutive trials on both tasks before nine trials, then learning was terminated, provided that the minimum of six trials had been completed. This meant that students in the spelling-present condition heard, spoke, or saw each word between 47 and 68 times, and students in the spelling-absent condition heard or spoke each word 53 to 77 times, depending on the number of learning trials. The order of treatment condition (written words present or absent) and the word set used in each treatment (Set 1 or Set 2) were counterbalanced. Children were randomly assigned to condition, with 5 children receiving each possible order and set.

### *Vocabulary Posttests*

Memory for the newly learned words was tested after a 1-day delay via three experimenter-devised tasks given in the order listed. Internal consistency reliabilities were estimated from our sample. In the word production task, students listened to the core definition of each word and recalled its pronunciation. The reliabilities were .56 in both the spelling and the no-spelling conditions. In the spelling production task, students heard each word and wrote its spelling. The reliabilities were .57 in the spelling condition and .60 in the no-spelling condition. In the recognition matching task, students were shown the six vocabulary words written across the top of a page, followed by 12 meaning-elaborated sentences that were heard during the word-learning task. Target words were replaced by blank spaces. The words and sentences were read aloud by the experimenter, and children chose the words that best completed the sentences. The reliabilities were .70 in the spelling condition and .41 in the no-spelling condition.

All testing was conducted with individual children over 3 consecutive days in 40-min sessions. On Day 1, children were taught one set of words in one of the conditions. On Day 2, they were posttested on their memory for Day 1 words and meanings and taught the second set of words in the other condition. On Day 3, they were tested on their memory for Day 2 words and then administered the tests of vocabulary and literacy.

## Results and Discussion

Performance on the literacy tests revealed that, on average, the second graders were reading at grade level on the WRMT-R Word Identification subtest ( $M = 2.2$  grade-equivalent score,  $SD = 0.68$ ). Also, they read and spelled about half of the CVC nonwords

correctly ( $M = 55\%$  read,  $M = 53\%$  spelled). However, their performance on the PPVT vocabulary test was substantially below national norms, with a mean standard score of 87.5 ( $SD = 11.0$ ). Scores of the Latin American students were very similar to scores of the African American students. These findings indicate that second graders in the present study possessed average word reading skills but below average vocabularies.

A preliminary analysis of vocabulary learning was conducted to evaluate the effects of the two control variables resulting from the counterbalanced design: word set (i.e., Set 1 vs. 2) and learning order (i.e., words with spellings learned before vs. after the words without spellings). The dependent measure was the difference between the number of trials to reach criterion in learning words with spellings and the number without spellings (i.e., criterion equaled three perfect successive trials). Two-way analyses of variance (ANOVAs) were conducted with word set and learning order as the between-subjects, independent variables. In one ANOVA, the dependent measure was the difference in pronunciation recall in the with- versus without-spellings conditions. In the second ANOVA, the dependent measure was the difference in definition recall in the with- versus without-spellings conditions. No significant main effects or interactions involving word set or task order were detected in either analysis (all  $ps > .05$ ). Hence, these control variables were dropped from further analyses. These ANOVAs also included a test of the intercept term, which is conducted as part of an ANOVA by the SPSS statistical program. When the dependent measure is a difference score, the intercept test indicates whether the mean difference is significantly different from zero. In the present analyses, the intercept tests were significant, favoring the spelling over the no-spelling condition,  $F(1, 16) = 8.53$ ,  $p < .01$ , for pronunciation recall;  $F(1, 16) = 18.44$ ,  $p < .01$ , for definition recall.

The next analyses examined the number of trials to criterion that students required to learn pronunciations and definitions. The independent variable was a repeated measure comparing the presence versus absence of spelling aids. Mean performance and test statistics are reported in Table 2. Paired-sample  $t$  tests revealed that learning of both pronunciations and meanings was significantly better with than without spellings. On the basis of Cohen's (1988) rule of thumb (i.e., .20 is a small effect, .50 is a moderate effect, and .80 is a large effect), effect sizes were moderate to large (see Table 2). The percentages of students reaching criterion in the two treatment conditions were quite disparate as well and favored spelling aids over no aids (see Table 2). These findings show that exposure to written words exerted a powerful effect on learning the pronunciations of new vocabulary words and their definitions.

To compare growth from trial to trial in the two tasks and conditions, we conducted an ANOVA with learning task (pronunciations vs. definitions), spelling condition (present vs. absent), and trials (Trials 1 through 5) as the independent variables. All were repeated measures. The dependent variable was correct recall. Results revealed significant main effects of task,  $F(1, 19) = 217.16$ ,  $p < .01$ ; spellings,  $F(1, 19) = 23.70$ ,  $p < .01$ ; and trials,  $F(4, 16) = 41.01$ ,  $p < .01$ , as well as significant interactions between task and spelling condition,  $F(1, 19) = 12.64$ ,  $p < .01$ , and between task and trials,  $F(4, 16) = 13.47$ ,  $p < .01$ . The remaining interactions were not significant (both  $ps > .05$ ). Mean performance is shown in Figure 1.



Table 2  
*Mean Performance Recalling Spoken Words and Meanings as a Function of Whether Students Were Exposed to the Spellings of Words in Experiment 1*

Tasks and measures	Spelling ( <i>n</i> = 20)	No spelling ( <i>n</i> = 20)	<i>t</i> (19)	Effect size <sup>a</sup> ( <i>d</i> )
Learning spoken words				
Trials to criterion (9 max)	7.9 (1.6)	8.7 (1.0)	3.11**	0.61
% reaching criterion	50%	15%		
Learning definitions				
Trials to criterion (9 max)	5.0 (1.8)	7.0 (2.3)	4.12**	0.98
% reaching criterion	100%	60%		
Posttests				
Recalling spoken words (6 max)	4.0 (1.5)	2.9 (1.5)	3.93**	0.73
Recalling spelling of words (6 max)	4.5 (1.5)	3.0 (1.7)	3.81**	0.94
Matching words to sentences (12 max)	11.5 (0.9)	10.8 (1.7)	1.93	0.54

Note. Standard deviations are in parentheses.

<sup>a</sup> Calculation of effect size: difference between means divided by pooled standard deviation.

\*\*  $p < .01$ .

Regarding the main effect of task, as is apparent in Figure 1, word definitions were recalled more easily than pronunciations of words across trials. Regarding the interaction between task and trials, differences favoring definitions over pronunciations were huge during earlier trials and diminished as memory for the pronunciations improved. Regarding the main effect of spelling condition, as seen in Figure 1, both pronunciations and definitions were learned more quickly with spelling aids than without aids. Regarding the interaction between task and spelling condition, the advantage of seeing spellings over not seeing them was greater when students learned pronunciations of words than when they learned definitions. From these findings, we conclude that the presence of written words during study and feedback periods helped children learn new vocabulary words and their definitions. Moreover, remembering pronunciations of the new words was harder than remembering their definitions.

To determine whether the benefit of spelling exposure held across words as well as students, we calculated the mean number of correct responses over the first five trials for each student learning each word and averaged these values. Comparison of means shown in Table 1 reveals that 100% of the spoken words

and 83% of the definitions showed the pattern favoring superior learning with spelling aids compared to no spelling aids. These findings confirm that the benefit of seeing spellings generalized across both words and students, and it held for learning meanings and pronunciations.

We gave posttests 1 day later to assess whether students remembered the spoken words and meanings after a delay. Mean performance and test statistics are reported in Table 2. Paired-sample *t* tests revealed that students recalled the spoken words significantly better in the spelling than in the no-spelling condition ( $M = 67\%$  vs.  $48\%$  correct). Also, students wrote the words more accurately if they had seen spellings than if they had not ( $M = 75\%$  vs.  $50\%$  correct), which indicates that the spellings had been stored in memory. These findings show that the benefit of spellings for vocabulary learning persisted at least 1 day later. On the measure of performance matching words to their sentences, the difference favoring spellings over no spellings fell short of statistical significance ( $p > .05$ ), possibly because of ceiling effects reflecting almost perfect recognition of word meanings ( $M = 90\%$  to  $96\%$  correct).

We calculated correlation coefficients to determine whether students' performance in our laboratory word learning tasks was related to their vocabulary and literacy skills, as assessed by tests of word and nonword reading, nonword spelling, and PPVT receptive vocabulary. The laboratory measures of vocabulary learning were the mean numbers of pronunciations and definitions correctly recalled on Trials 1–5 in the spelling and no-spelling conditions. Findings revealed that the WRMT–R word reading test correlated strongly with recall of pronunciations in the spelling-present condition ( $r = .67, p < .01$ ) and the no-spelling condition ( $r = .55, p < .05$ ). The nonword reading test showed the same pattern ( $r = .49, p < .05$ , and  $r = .47, p < .05$ , respectively). The PPVT vocabulary test correlated significantly with recall of definitions in the spelling-present condition ( $r = .48, p < .05$ ) but fell short of significance in the spelling-absent condition ( $r = .40, .05 < p < .10$ ). None of the other correlations reached statistical significance (all  $ps > .05$ ). Findings suggest that the ability to remember the pronunciations of new vocabulary words, whether

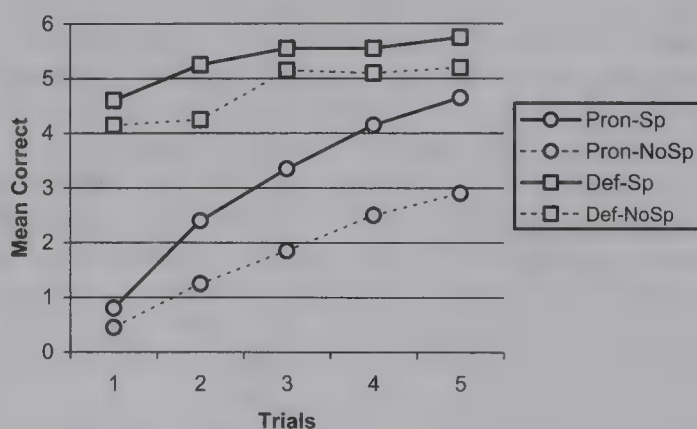


Figure 1. Mean number of pronunciations recalled in the spelling condition (Pron-Sp) and in the no-spelling condition (Pron-NoSp), and mean number of definitions recalled in the spelling condition (Def-Sp) and in the no-spelling condition (Def-NoSp) over five learning trials in Experiment 1.

learned with or without spelling aids, is influenced by students' orthographic knowledge.

## EXPERIMENT 2: VOCABULARY LEARNING WITH FIFTH GRADERS

Experiment 2 was conducted to replicate and extend results of the first experiment. The participants were fifth rather than second graders. Ten multisyllabic words, rather than six monosyllable words, were taught in each condition (i.e., with vs. without spelling aids). The sample size was enlarged from 20 to 32 participants. Posttests were expanded to include transfer tasks.

One question of interest was whether exposure to spellings would help older students learn pronunciations and meanings of the words better than no exposure. Because older students are more proficient decoders and possess greater knowledge of written and spoken words, they may anticipate spontaneously how unfamiliar spoken words are spelled when they hear and practice them, even without spelling aids (Stuart & Coltheart, 1988). If so, exposure to spellings may exert a smaller influence on their vocabulary learning than on younger students' learning. Alternatively, the opposite effect might occur. According to Ehri and Wilce (1979), having greater knowledge of the orthographic system should facilitate memory for written words, particularly if the words are multisyllabic and, hence, harder to remember.

### Method

#### Participants

The participants were 32 fifth graders from the same school as those in Experiment 1. The sample included 31 African Americans, 1 Latin American, 18 girls, and 14 boys whose parents had given written consent to participate. Mean chronological age was 10 years 11 months. Participants were tested in the spring. Teachers verified that no one exhibited any cognitive or sensory impairment or received English language services.

#### Materials and Procedures

##### Literacy Assessment

Several tests were given. The purpose was to assess students' vocabulary and literacy skills.

*Reading words.* Graded lists, each of which had 20 words, from second through eighth grade, were taken from the Boder Test of Word Reading (Boder & Jarrico, 1982). Students were allowed 5 s to read each word. Reading was stopped when students were unable to read more than 6 words in a list. The split-half reliability in our sample was .98. Conversion of scores to grade-equivalent values was based on list-level performance rather than national norms, which are not provided in the test manual.

*Test of Word Reading Efficiency (TOWRE) Test of Phonemic Decoding Efficiency (Torgesen, Wagner, & Rashotte, 1999).* Children were given 45 s to read as many nonwords as possible from a list of increasingly difficult items. The score was based on the number pronounced correctly. According to the manual, test-retest reliability is .83.

*Ganske Spelling Inventory (Ganske, 2000).* Twenty progressively more difficult words were pronounced and embedded in

sentences. Children repeated and wrote the words. The task was terminated if fewer than two words in a set of five were correct. Scores were based on the number correctly spelled. The split-half reliability in our sample was .90.

*PPVT-III.* Students were administered the PPVT-III.

##### Vocabulary Learning

Two sets of 10 concrete, multisyllabic, low-frequency nouns were taught. The same rare word Web sites were consulted to locate the words as in Study 1. Teachers judged that their students would not know them. Also, most of the words proved unfamiliar to a sample of graduate students. The words were rated as occurring fewer than four times per million words, with 16 of the words so infrequent that they were not rated (Thorndike & Lorge, 1972).

Words within each of the two sets began with a different consonant. Each set consisted of 5 two-syllable and 5 three-syllable words. The words and core definitions appear in Table 3. Three definitions were simplified or made concrete to be more readily understood, more easily depicted, or more appropriate for fifth-grade learners (i.e., *koomkie*, meaning a female elephant in heat; *frenulum*, meaning the membrane that holds the tongue back from extending too far; and *juggernaut*, meaning a large, overpowering force or object; see Table 3 for the simplified definitions that were taught). Four meaning-clarifying sentences were constructed for each word. The format of the materials and the procedures used to teach and test recall of the words and meanings were the same as in Experiment 1, except that students were given a minimum of five and a maximum of eight trials to learn the words and their meanings. This meant that in the spelling-present condition, students heard, spoke, or saw each word 40 to 61 times, and students in the spelling-absent condition heard or spoke each word 45 to 69 times, depending on the number of learning trials. The order of the treatment (i.e., whether students learned words with or without spellings first) and the word set taught in each treatment were counterbalanced. Children were randomly assigned to condition, with 8 children receiving each possible order and set.

##### Vocabulary Posttests

Memory for spoken words and meanings was tested after a 1-day delay with five experimenter-devised tasks that were given in the following order. Internal consistency reliabilities were estimated.

*Oral cloze task.* In this transfer task, students' ability to recall the spoken forms of target words and fit them in appropriate sentence contexts was assessed. The experimenter read orally 20 new sentences that included the words and clearly expressed their meanings. Students listened to each sentence with the target word omitted and pronounced the missing word. The reliabilities were .67 in the spelling condition and .60 in the no-spelling condition.

*Recall spoken words.* Students were shown the same pictures used during training but without any written words and recalled the words. The reliabilities were .72 in the spelling condition and .67 in the no-spelling condition.

*Recall definitions of words.* Students heard each word pronounced and provided its definition. No written words were present. Students performed almost perfectly, making it impossible to estimate this test's reliability with our data.



Table 3

*Mean Number of Fifth Graders Who Recalled Each Spoken Word Correctly Per Trial During the First Five Learning Trials and Who Recalled Each Meaning Correctly Per Trial During the First Three Learning Trials as a Function of Whether Students Were Exposed to Spellings of the Words in Experiment 2*

Vocabulary words and meanings	Recall spoken		Recall meaning	
	Sp.	No sp.	Sp.	No sp.
Barrow—a small hill	13.2	10.6	15.7	12.7
Juggernaut—a large truck	11.0	6.2	14.7 <sup>a</sup>	16.0 <sup>a</sup>
Wimple—what nuns wear on their heads	10.6	6.8	12.3 <sup>a</sup>	13.7 <sup>a</sup>
Mullock—a pile of trash	9.8	7.0	13.0	11.0
Tandem—a horse-drawn carriage	9.8	3.8	14.0	12.7
Fribble—a foolish, shallow person	9.8	2.8	13.0	11.0
Potlatch—a Native American festival where the chief gives out gifts	9.6	9.4	14.0	13.0
Dobson—a bug that is a kind of beetle	9.4	6.2	15.0	11.0
Muleta—a red cape	9.2	5.2	15.0	13.0
Koomkie—a female elephant that attracts male elephants	9.2	4.4	15.7	13.7
Vibrissa—the whiskers on a cat	9.2	3.4	15.3	14.7
Gangrel—a homeless person	8.8	8.6	15.3 <sup>a</sup>	15.3 <sup>a</sup>
Tamarack—a big tree found all over America	8.0	2.2	13.3	10.3
Chigger—a kind of bug that eats the blood of animals and people	7.8	4.4	14.3	13.7
Hicatee—a kind of turtle that lives in the water	7.4	5.6	14.0	12.7
Laburnum—a small tree with bright yellow flowers	6.8	3.0	14.0	11.3
Kerfuffle—a fuss or fight	6.2	4.0	14.7	12.0
Proboscis—a really big nose	6.0	5.2	12.7	10.7
Frenulum—a tongue	5.8	2.0	14.7	10.7
Scrivello—the tusks on an elephant	4.4	2.2	12.7	9.0

*Note.* The maximum number of students who recalled a word correctly was 16. Sp. = spelling.

<sup>a</sup> There were three exceptions to the general pattern of superior learning when spellings were seen compared with when spellings were not seen, all involving the recall of word meanings.

*Recognize meanings of words.* In this transfer task, students were shown four new pictures not previously seen during training, a target word was spoken by the tester, and students selected the picture that best illustrated the word. The foils consisted of unseen pictures of other target words. Each picture appeared multiple times, once as a target, and other times as a foil. Students performed almost perfectly, making it impossible to estimate this test's reliability with our data.

*Spell words.* Students heard each target word pronounced, repeated it, and wrote its spelling. The reliabilities were .81 in the spelling condition and .61 in the no-spelling condition.

All testing was conducted with individual children over 3 consecutive days and followed the same schedule used in Experiment 1 (see above). Each session lasted about 40 min.

## Results and Discussion

Performance on the Boder Test of Word Reading revealed a bimodal distribution, which provided the basis for dividing students by orthographic knowledge into a lower level group ( $n = 18$ ), who read fewer than 93 words, and a higher level group ( $n = 14$ ), who read more than 106 words. The lower group's mean performance was below grade level, whereas the higher group's mean performance was above grade level (see Table 4). Characteristics and mean performance of the two groups on the other tests are reported in Table 4. Results of  $t$  tests revealed that the groups did not differ significantly in receptive vocabulary, which was

somewhat below the national average, but they clearly differed in their nonword decoding and spelling skills. Correlation coefficients revealed strong relationships among the measures of word and nonword reading and spelling, with correlations ranging from .69 to .82 ( $ps < .01$ ). These findings show that higher and lower level readers differed in their knowledge of the orthographic system but not in their receptive vocabulary knowledge.

We conducted a preliminary analysis of vocabulary learning to evaluate the effects of two control variables resulting from the counterbalanced design: word set (i.e., Set 1 vs. 2) and learning order (i.e., spelling present condition completed first vs. second). The dependent measure was the difference between the number of trials to criterion to learn words with spellings and the number of trials to learn words without spellings (i.e., criterion equaled three perfect successive trials). Two-way ANOVAs were conducted with word set and learning order as the between-subjects, independent variables. No significant main effects or interactions involving word set or task order were detected (all  $ps > .05$ ). Hence, these control variables were dropped from further analyses. However, results revealed a significant effect of intercept, indicating that the mean difference score was significantly greater than zero, favoring the spelling over the no-spelling condition,  $F(1, 28) = 19.55$ ,  $p < .01$ , for word recall;  $F(1, 28) = 33.56$ ,  $p < .01$ , for definition recall.

The next two analyses examined the numbers of trials to criterion that students required to learn pronunciations and definitions.

Table 4  
*Student Characteristics and Mean Performance on Literacy and Language Tests as a Function of Reader Level as Determined by the Boder Word Identification Test in Experiment 2*

Characteristics and measures	Reader level		<i>t</i> (30)
	Lower ( <i>n</i> = 18)	Higher ( <i>n</i> = 14)	
Gender (girls; boys)	9; 9	9; 5	
Ethnicity			
African American	17	14	
Latin American	1	0	
Age (years; months)	10; 11	10; 10	
Boder Word Identification (140 max)	71.7 (15.3)	125.1 (7.0)	13.12 <sup>a,**</sup>
Approximate grade equivalent	4.6	7.3	
TOWRE Nonword Decoding Efficiency			
Raw score (63 max)	16.72 (10.3)	33.57 (12.7)	4.13 <sup>**</sup>
Grade equivalent score	2.2	4.8	
Ganske Spelling Inventory (20 max)	7.9 (3.1)	13.1 (3.1)	4.62 <sup>**</sup>
PPVT Receptive Vocabulary standard score	89.1 (9.2)	92.9 (12.0)	0.99

*Note.* Standard deviations are in parentheses. TOWRE = Test of Word Reading Efficiency; PPVT = Peabody Picture Vocabulary Test.

<sup>a</sup> Equal variance not assumed; *t*(24.92).

<sup>\*\*</sup> *p* < .01.

The independent variables were spelling condition (present vs. absent, within subject) and reader level (higher vs. lower, between subjects). Means and test statistics are reported in Table 5. In the ANOVA of spoken word recall, significant main effects of spelling condition and reader level and a significant interaction between the two variables were detected. Means in Table 5 reveal that word learning was superior in the spelling-present condition. Higher level readers outperformed lower level readers, and the advantage provided by seeing written words over not seeing written words was much greater among higher readers than among lower readers. The difference in percentages of readers who reached criterion in learning pronunciations with spellings compared to no spellings was much greater among the higher readers (i.e., 79% vs. 29%, for a difference of 50%) than among lower readers (i.e., 28% vs. 0%, for a difference of 28%). Whereas the majority of higher readers reached criterion in learning pronunciations with spellings, only a minority of lower readers reached criterion with spellings. When spellings were not seen, some higher readers (29%) but no lower readers reached criterion in recalling pronunciations. These findings show that giving students written as well as spoken vocabulary words to study exerted a strong impact on their memory for the pronunciations of new vocabulary words, especially students with stronger orthographic knowledge.

In the ANOVA of trials to criterion recalling definitions, significant main effects of spelling condition and reader level were detected, but there was no significant interaction. From mean performance in Table 5, it is apparent that definitions were learned more readily when written words had been seen than when they had not been seen. Also, higher level readers recalled definitions better than lower level readers. These findings show that exposure to spellings helped both higher and lower readers learn the meanings of words. Comparison of performance recalling spoken words and definitions in Table 5 shows that it was easier for students to learn definitions than to remember pronunciations of the words. Most of the higher and lower readers reached criterion in learning

definitions, but many did not reach criterion in recalling pronunciations of the words.

To examine performance over trials, we conducted three-way ANOVAs. The independent variables were spelling condition, reader level, and trials (Trials 1–5). The dependent variables were the number of correct pronunciations recalled in one ANOVA and the number of correct definitions recalled in the second ANOVA. For pronunciation recall, significant main effects of spelling condition and reader level and a significant interaction between these two variables were detected (all *ps* < .01). Performance patterns favored spellings-present over spellings-absent conditions, higher over lower level readers, and a greater advantage of spellings over no spellings among higher readers than among lower readers. Mean performance is displayed in Figure 2. These results duplicate those on the trials-to-criterion measure reported in Table 5.

Of particular interest in this analysis were effects involving trials showing a significant main effect,  $F(4, 27) = 117.19$ ,  $p < .01$ , as well as significant interactions with reader level,  $F(4, 27) = 2.71$ ,  $p = .05$ ; with spelling exposure,  $F(4, 27) = 14.77$ ,  $p < .01$ ; and with Reader  $\times$  Spelling exposure,  $F(4, 27) = 3.81$ ,  $p < .05$ . From Figure 2, it is apparent that spoken word recall improved across trials. Both groups showed better recall of pronunciations with than without spellings after Trial 1. Gains from Trials 1 to 3 were much greater for higher readers learning words with spelling aids than gains in the other three conditions (i.e., higher level readers learning words without spellings, and lower level readers learning words both with and without spellings). After Trial 3, ceiling effects suppressed continued gains of the higher group with spelling aids. In other words, higher level readers derived an increasing advantage over lower level readers from spelling aids in recalling pronunciations as word learning proceeded. These findings reveal a pattern of the rich getting richer as a result of superior orthographic knowledge when spellings were provided during vocabulary learning.



Table 5  
Mean Performance on the Vocabulary Learning Tasks and Posttests as a Function of Reader Level and Spelling Condition in Experiment 2

Tasks and measures and reader levels	Spelling <i>M</i> ( <i>SD</i> )	No spelling <i>M</i> ( <i>SD</i> )	Effect size <sup>a</sup>	<i>F</i> (1, 30) <sup>b</sup>
Vocabulary learning task				
Recall spoken words				
Trials to criterion (8 max)	7.1 (1.2)	7.9 (0.2)	-1.14	28.88** (S)
Higher level readers	6.4 (1.4)	7.9 (0.4)	-1.67	11.77** (R)
Lower level readers	7.7 (0.7)	8.0 (0)	-0.86	11.69** (S × R)
% reaching criterion				
Higher level readers	79	29		
Lower level readers	28	0		
Recall definitions				
Trials to criterion (8 max)	4.6 (1.3)	6.2 (1.3)	-1.23	33.14** (S)
Higher level readers	4.1 (0.9)	5.7 (1.5)	-1.33	6.99* (R)
Lower level readers	5.0 (1.5)	6.6 (1.0)	-1.23	<1 (S × R)
% reaching criterion				
Higher level readers	100	93		
Lower level readers	94	89		
Posttests				
Recall spoken words (10 max)	7.3 (2.3)	5.3 (2.4)	0.83	20.32** (S)
Higher level readers	8.8 (0.8)	6.0 (2.6)	1.65	9.87** (R)
Lower level readers	6.1 (2.4)	4.8 (2.1)	0.57	2.80 (S × R)
Spell words (10 max)	4.9 (2.8)	1.9 (1.8)	1.30	96.32** (S)
Higher level readers	7.1 (1.6)	2.9 (1.8)	2.47	22.82** (R)
Lower level readers	3.3 (2.4)	1.2 (1.4)	1.11	10.65** (S × R)
Embed words in cloze (10 max)	6.6 (2.4)	4.7 (2.3)	0.79	21.99** (S)
Higher level readers	8.2 (1.2)	5.2 (2.7)	1.50	9.31** (R)
Lower level readers	5.4 (2.4)	4.2 (2.0)	0.55	4.26* (S × R)

Note. There were 14 higher level readers and 18 lower level readers.

<sup>a</sup> Calculation of effect size: difference between means divided by pooled standard deviation.

<sup>b</sup> *F* values in analyses of variance for main effects of spelling condition (S), reader level (R), and interaction between condition and level (S × R).

\*  $p < .05$ . \*\*  $p < .01$ .

In the ANOVA of definition recall over trials, significant main effects of spelling condition and reader level were detected. Performance patterns favored spellings present over spellings absent ( $p < .01$ ) and higher over lower level readers ( $p < .05$ ). These results duplicate those on the trials-to-criterion measure reported in

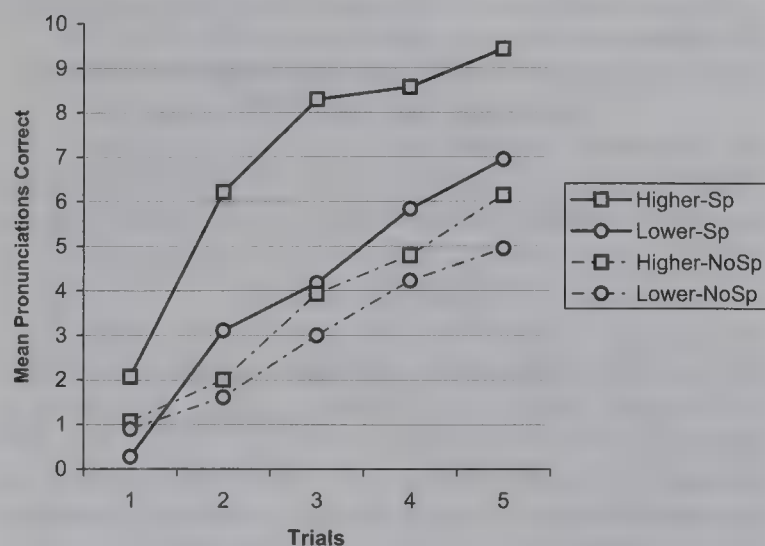


Figure 2. Mean number of pronunciations recalled over five learning trials by higher level readers and by lower level readers in the spelling condition (Sp) and the no-spelling condition (NoSp) in Experiment 2.

Table 5. A significant main effect of trials was detected,  $F(4, 27) = 28.19$ ,  $p < .01$ , but there were no significant interactions between trials and the other variables (all  $ps > .05$ ). As shown in Figure 3, mean performance improved from Trials 1 through 3 in all conditions, but after that performance was close to ceiling. Both

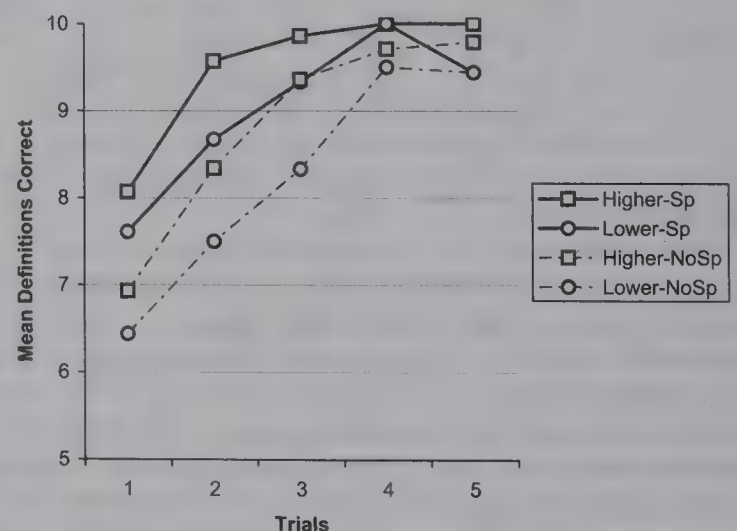


Figure 3. Mean number of definitions recalled over five learning trials by higher level readers and by lower level readers in the spelling condition (Sp) and the no-spelling condition (NoSp) in Experiment 2.

higher and lower readers displayed superior definition recall in the spelling than the no-spelling condition from the very first test trial.

In summary, these findings show that fifth graders benefited from spelling aids in learning new vocabulary words. Effect sizes ranged from moderate to very large (see Table 5). Spoken words and their meanings were learned faster when spellings were seen during study and feedback periods than when they were not. In addition, the higher level readers learned pronunciations and definitions more easily than the lower level readers, and the higher level readers benefited much more from spellings compared to no spellings in recalling spoken words than the lower level readers.

To determine the extent that spelling aids boosted the learning of individual words, we calculated the mean number of correct responses over the first five trials for each student learning each word and averaged these values. Results are presented in Table 3. Comparison of mean recall values revealed that 100% of the spoken words and 80% of the definitions benefited from spelling exposure. These findings show that the advantage of spelling aids for learning both pronunciations and meanings generalized across words as well as students.

Half of the words that students learned contained two syllables, and half were three syllables. Gathercole and Baddeley (1989) found that learning was more difficult when words contained more than two syllables. Table 3 shows the words ordered from easy to difficult on the measure of spoken word recall in the spelling-present condition. Most of the three-syllable words (80%) are found in the bottom 10, harder to learn words. In contrast, most of the two-syllable words (80%) sit among the top 10, easier to learn words. Correlations between the number of syllables and recall values revealed significant negative relationships for recall of pronunciations with spelling aids ( $r = -.59, p < .01$ ) and recall of pronunciations without spelling aids ( $r = -.53, p < .05$ ), but relationships were not significant for recall of definitions ( $r = -.06, p > .05$ , for the spelling condition, and  $r = -.21, p > .05$ , for the no-spelling condition). Clearly, syllable length influenced recall of pronunciations but not of definitions.

Analyses of the errors characterizing students' responses over trials in the word-recall task revealed that learning to pronounce words correctly was the central difficulty. Producing words but matching them to the wrong pictures accounted for only 1% to 2% of the errors. Most errors involved either a failure to respond or a failed attempt at producing the correct pronunciation. In the latter case, mispronunciations bore a closer resemblance to correct pronunciations when spellings had been seen than when they had not been seen. Comparison of the proportion of mispronunciations that shared any syllables or initial sounds with the correct words revealed a significant difference favoring those in the spelling-seen condition ( $M = 30\%$  seen vs.  $17\%$  not seen),  $t(31) = 4.23, p < .01$ . This indicates that spellings strengthened students' memory for partial sounds in the words being learned.

Several posttests were given to assess longer term effects of vocabulary learning, including students' memory for the spoken words depicted in training pictures, their ability to fit the new vocabulary words into cloze sentences, and their memory for the spellings of words. ANOVAs were conducted with spelling condition and reader level as the independent variables. Mean performance and test statistics in Table 5 show that on all of these measures, main effects of reader level and spelling condition were statistically significant. The interaction was significant on the

cloze test and spelling test but not on the spoken-word recall test. As apparent in Table 5, students recalled spoken words, wrote their correct spellings, and embedded words in sentences significantly better when they had learned the words by seeing their spellings than when they had not seen their spellings. These findings show that spelling effects persisted for at least a day beyond the end of training.

Higher level readers outperformed lower level readers on all three posttests, as shown in Table 5. On the spelling and cloze measures, significant interactions resulted from higher level readers benefiting more from spellings than no spellings compared to lower readers. The percentage of words spelled correctly increased from 29% without spellings to 71% with spellings for higher readers but only from 12% without spellings to 33% with spellings for lower readers. The fact that higher level readers had strong memory for the spellings of words they had seen during the word-learning task supports the idea that memory for spellings was the reason that higher readers learned words better when they were exposed to spellings than when they were not.

Posttests also assessed students' memory for the definitions of words and their ability to recognize which of four previously unseen pictures depicted each vocabulary word. Performance of both higher and lower readers was virtually perfect in both of these tasks, with means of 99% to 100%. Thus, students had no difficulty remembering or recognizing the meanings of the new vocabulary words 1 day later regardless of whether spellings had been seen.

Correlations between the literacy and language measures (i.e., Boder word reading, Ganske word spelling, TOWRE nonword reading, PPVT vocabulary recognition) and performance on the vocabulary learning tasks were examined. In the learning tasks, the measure of spoken word recall was the mean performance on Trials 1 through 5. The measure of definition recall was the number of trials to criterion, because ceiling effects suppressed individual differences on the measure with one to five trials. The literacy measures (i.e., word and nonword reading and spelling) were highly related to each other (correlations ranging from .69 to .82), indicating that there was much overlap in the skills assessed. Reported below are the statistically significant correlations at  $p < .05$ . Note that negative correlations associated with recalling definitions resulted from the trials-to-criterion measure, with lower scores reflecting faster learning.

For learning with spelling aids,

1. recalling pronunciations: .71 (reading words), .66 (spelling words), and .56 (reading nonwords);
2. recalling definitions:  $-.41$  (reading words) and  $-.41$  (spelling words).

For learning without spelling aids,

1. recalling pronunciations: .59 (spelling words) and .45 (reading nonwords);
2. recalling definitions:  $-.41$  (reading words),  $-.67$  (spelling words), and  $-.51$  (reading nonwords).

These results reveal the heavy influence of word reading and spelling skills in all of the vocabulary learning tasks. At least two



possible explanations can be identified for the strong relationship between spelling ability and recall of pronunciations and meanings when spellings were not seen. The better spellers might have been more strategic and spontaneously imagined spellings to help them store the words in memory. Alternatively, better spellers may have richer, more elaborated phonological memories as a result of acquiring more extensive knowledge of mapping relations between graphemes and phonemes. In contrast to Experiment 1, which revealed a significant correlation for second graders, scores on the PPVT-III vocabulary test were not significantly correlated with the definition learning measures for fifth graders (i.e.,  $r_s = -.03$  and  $-.15$ ,  $p_s > .05$ ). Rather, word reading and spelling abilities were more related to definition scores than was receptive vocabulary knowledge. Why this occurred is not clear. PPVT-III scores were not correlated with word reading or spelling scores either (i.e., correlations ranged from  $-.02$  to  $.16$ ,  $p_s > .05$ ).

### GENERAL DISCUSSION

In summary, the results of both experiments yield several important findings. Elementary school students learned and remembered the pronunciations and meanings of new vocabulary words better when they were exposed to written forms of the words during study periods than when they only heard and repeated the words. The advantage was evident on delayed posttests assessing transfer as well as training effects. Spelling aids proved advantageous for learning most, if not all, of the individual vocabulary words' pronunciations and meanings, suggesting that the effect applies to the learning of words in general. The advantage of spelling aids was evident among fifth graders as well as second graders, indicating that the benefit was not limited developmentally. Higher level fifth-grade readers, who had superior orthographic knowledge, benefited more from spellings than lower level fifth-grade readers. In fact, higher level readers made increasingly greater gains from one trial to the next with spelling aids compared to gains with no spelling aids and compared to the gains made by lower level readers. Learning and remembering the pronunciations of new vocabulary words proved harder than learning definitions of the words, so exposure to spellings helped students with the more difficult part of vocabulary learning.

The advantage provided by spelling exposure in learning pronunciations and meanings of new vocabulary words can be explained by Ehri's (1992) theory of word learning. According to the theory, word learning entails the amalgamation of various word identities in memory, including orthographic, phonological, and semantic identities. When readers hear an unfamiliar word and comprehend its meaning, an amalgamated representation combining phonological and semantic identities begins to form in memory. If readers also view the word's written form, connections are activated between graphemes in the spelling and phonemes in the pronunciation. These connections bond the spelling to its pronunciation, and the word's orthographic identity is added to its representation in memory. Because pronunciations of words are fleeting, graphemes in spellings serve to clarify phonemic constituents in memory. Written words also represent pronunciations in another modality as a visual image that enhances the word's memorability, according to dual coding theory (Sadoski & Paivio, 2001). Thus, when spellings of new words are seen and pronunciations are heard, students are able to form stronger, better specified phono-

logical representations in memory earlier during the learning process than when spellings are not seen. Earlier formation of pronunciations in memory speeds up the incorporation of semantic identities into the representations as well.

In the present study, several findings were consistent with this theory. When students misrecalled pronunciations of words, they preserved more correct sounds in the spelling exposure condition than in the no-exposure condition. This supports the idea that spellings became connected to pronunciations and enhanced memory for their parts during learning. Students began to recall pronunciations and meanings of words earlier during the learning trials when they saw spellings than when they did not. This supports the idea that spelling-based amalgams began to form in memory sooner than amalgams without spellings. Posttest performance occurring 1 day after learning revealed that students wrote vocabulary words more accurately when they had seen them than when they had not, indicating that spellings of specific words were stored in memory.

Several alternative explanations for facilitation can be dismissed. It was not the case that students in the spelling aids condition were exposed to words more frequently. A tally revealed that, in fact, exposure was greater in the spelling-absent condition. It is not likely that seeing spellings caused children to repeat words more accurately, thus improving the quality of their phonological representations in memory. The words contained familiar sounds, syllables, and blends. Children had no trouble repeating them correctly and easily in the spelling-absent condition, and, in fact, they repeated them more frequently than in the spelling-present condition.

It is not likely that seeing spellings induced students to apply a decoding strategy to sound out and blend the words, thus giving them extra practice. Our procedures limited the need for this because the experimenter spoke the words as soon as they were shown. Also, the written words were visible for a matter of seconds, thus limiting rehearsal time. Procedures were, in fact, structured to give more practice to students in the no-spelling condition. They heard and pronounced the words more times than students in the spelling condition. No child who viewed written words was observed to overtly decode them. It is more likely and consistent with our preferred explanation that these children engaged in a different form of decoding. When they saw spellings and heard or spoke pronunciations, they automatically computed the connections between graphemes and phonemes to determine how spellings mapped onto pronunciations. It is this orthographic mapping process that, in our view, explains what students did when they saw and heard written words and accounts for the effect of seeing spelling on vocabulary learning.

Another reaction to our findings might be to dismiss them as patently obvious because two modalities should always be better than one. However, in the case of vocabulary learning, it is not clear that combining visual and verbal modalities in the form of spellings and pronunciations should make vocabulary learning any better. First, the phonological working memory theory of Gathercole and Baddeley (1993) does not regard spellings of words as strengthening memory for their phonological representations. Neither have vocabulary researchers been explicit about how spellings contribute to vocabulary learning. In addition, older theories of the reading process have assigned a peripheral role to spellings in lexical memory. According to Goodman (1976), letter cues in



words are merely sampled during text reading rather than accessed in memory. According to Gough (1972), readers transform written words into blends of phonemes by applying grapheme-phoneme knowledge and then use the phonological forms of words to retrieve their meanings in memory. With practice, the transformation process speeds up, but the written forms of words are not mapped onto pronunciations in memory. According to some versions of dual route theory, written words are stored in memory, but as nonphonological forms linked to meanings (i.e., the direct, visual route; e.g., Barron, 1986). Thus, it is not obvious that two modalities should be better than one in the case of storing new vocabulary words in memory.

Observations of children's comments during the word learning trials bolster the claim that spellings themselves were stored and accessed to enhance memory for pronunciations. In the spelling-present condition of Experiment 1 during the word recall trials, 2 students remembered and orally named letters in the spellings of words before pronouncing them. A 3rd student was observed to cover her eyes while trying to pronounce the word, effectively trying to "see" the image of the word's spelling. Likewise, in Experiment 2, in the spelling aid condition during test trials when spellings were not present, 2 students orally recalled the names of letters in the words before attempting to pronounce them. Another student mispronounced the word on the test trial and then, on being shown the training card, said, "Oh, I misspelled it." A 4th student, when trying to recall the word *hicatee*, said, "I know there are two *es* at the end." These observations suggest that students had formed orthographic representations and attempted to use them to "read" pronunciations of words from memory.

Results of the present study replicate those of Ehri and Wilce (1979), who found that second graders learned the pronunciations of nonwords better with spelling aids than without them. Our findings extend their work by showing that spelling aids facilitated the learning of real words, including their meanings as well as pronunciations. Also, our findings show that spelling aids benefited vocabulary learning in older students. There was uncertainty about this. One reason why older students might not show much effect is that, compared to beginning readers, they possess more extensive decoding proficiency and phonological and orthographic knowledge of words. Conversely, when faced with the task of learning vocabulary words that are multisyllabic and, hence, harder to remember than CVC words, older readers might show an even larger effect than younger readers.

Findings support the latter hypothesis. Although different words were learned by the older and younger students, the words were not more difficult relative to the students' age and grade level. About the same percentage of students reached criterion in learning their set of words (i.e., 50% of second graders and 50% of fifth graders learned words with spellings, and 15% of second graders and 13% of fifth graders learned words without spellings). This comparability allows us to compare the effect sizes for older and younger students. On the measure of trials to criterion learning pronunciations of the words, the effect sizes favoring spelling aids over no spellings were  $-1.67$  (fifth-grade higher level readers),  $-0.86$  (fifth-grade lower level readers), and  $-0.61$  (second-grade readers). Thus, fifth graders showed bigger effects of spelling aids than second graders, especially fifth graders who were the higher level readers. These findings suggest two factors that magnify the benefit of spelling aids in learning new words: the number of

phonemes and syllables, and learners' knowledge of the orthographic system.

Other studies have reported word length effects on children's pronunciations of words. Gathercole and Baddeley (1989) found that nonwords longer than two syllables were harder for 4- and 5-year-olds to repeat than shorter words, and difficulty increased with word length. The present findings are consistent. Fifth graders had a harder time recalling the pronunciations of three-syllable words than two-syllable words during the learning trials, despite facilitation effects provided by spelling aids.

The essence of vocabulary learning is linking pronunciations to meanings of new words in memory. Gathercole and Baddeley (1993) provided theory and evidence explaining how phonological representations of new words are established in memory. A key process is phonological short-term memory, which maintains the word in working memory until a more permanent phonological representation is stored. Good readers are known to have better phonological working memories for words than poor readers, and this is thought to explain differences in vocabulary learning. The present findings, however, suggest that phonological memory may be less important than orthographic knowledge for explaining good-poor reader differences in learning the pronunciations of new vocabulary words when spellings of the words are seen. Comparison of the performance of higher and lower level readers in Figure 2 shows that the higher readers outperformed the lower readers only modestly and only on later trials in learning pronunciations from their spoken forms alone. However, higher readers were far superior to lower readers in learning pronunciations when they had the opportunity to amalgamate spellings to phonological representations of words in memory. This suggests that when students acquire strong orthographic skills and when new vocabulary words are read rather than just heard, orthographic processes lessen their dependence on phonological working memory for storing new vocabulary words in memory.

It is interesting that in the present study, we found that students' orthographic knowledge was strongly correlated with their memory for spoken words when spellings were not seen as well as when they were seen. This suggests that becoming literate in an alphabetic writing system facilitates phonological memory in general, not just when written words are seen. In a longitudinal study, Gathercole and Baddeley (1989) found that short-term working memory effects on vocabulary learning were diminished in children above 6 years of age. Whereas the correlation between vocabulary and nonword repetition scores ranged from .52 to .56 among children between the ages of 4 and 6 years, it declined to .28 at 8 years of age. The authors attributed the decline to older children's accumulation of more spoken words in memory and more practice repeating words (Gathercole, 2006). However, our findings suggest another explanation: that diminution results from the emergence of literacy skills, in particular the acquisition of orthographic knowledge and the linkage of spellings to the phonological forms of words in memory. This possibility invites further research.

The impact of orthography on phonological processing in tasks other than vocabulary learning has been documented (Ehri, 1984, 1985, 1993). Spellings of words have been found to influence how people segment words into phonemes and syllables, what sounds they conceptualize in words, and how quickly they distinguish spoken pairs of rhyming words that either share or do not share



spellings (e.g., *foam–roam* vs. *bomb–comb* vs. *goal–bowl*). However, such effects appear not to be generally recognized. Gathercole's (2006) work "Nonword Repetition and Word Learning: The Nature of the Relationship" appeared as a keynote article in a special issue of *Applied Psycholinguistics*. It was followed by 14 commentaries, none of which raised the possibility that orthographic representations of words might be important.

The present findings show the importance of students' acquisition of strong orthographic knowledge to benefit vocabulary learning. Fifth graders with better orthographic knowledge outperformed those with weaker knowledge on the training measures as well as on the posttests. Most impressive were the increasingly large gains over trials that higher level readers made when the pronunciations of words were learned with spellings compared to the gains made with no spellings and compared to the gains made by lower level readers (see Figure 2). The fact that higher level readers showed a much steeper slope in learning the pronunciations of words with spelling aids than lower level readers suggests a Matthew effect (Stanovich, 1986), with the rich getting richer in expressive vocabulary during the course of learning as a result of superior orthographic knowledge. Whereas a Matthew effect was observed in learning pronunciations, it was not observed in learning definitions, for which a ceiling effect might have suppressed differences.

Given these findings, it is perhaps surprising that higher readers did not show larger vocabularies than lower readers on the PPVT receptive vocabulary test. It may be that a Matthew effect was limited to expressive use of new vocabulary words that had been learned with spelling aids, or it may be that the PPVT-III test lacked sufficient sensitivity with this population of low-socioeconomic status, minority students who had below average vocabularies. Which is true awaits further study.

The fact that written words were processed and remembered at all during the learning trials might surprise some researchers, given that no attention was drawn to them and that the words did not have to be read because the experimenter pronounced them when they were shown. Also, pictures of the words were shown, and there is some research indicating that written words are overlooked when pictures accompany them (Sadoski, 2005). Our findings can be interpreted to bear on Share's (1995, 1999, 2004) self-teaching mechanism by raising a question about how it might operate. Share claimed that learners apply a decoding procedure to retain information about the spellings of words in memory. However, the present findings suggest that it is not the conscious application of a decoding strategy (i.e., sounding out letters and blending them) that is critical for memory but rather the implicit, spontaneous activation of mapping relations that connect graphemes to phonemes when spellings are seen and pronunciations are heard. A question for future research is whether students in the present study would have benefited even more from spelling aids if their attention had been directed to the spellings and if they had been instructed to decode the words aloud before the experimenter pronounced them.

Although the advantage was less than that for higher level readers, it is noteworthy that fifth-grade, lower level readers still benefited from spelling aids in their vocabulary learning. This benefit occurred despite the fact that their decoding proficiency on the TOWRE test was well below grade level (i.e., mean grade equivalent = 2.2). A smaller effect might have resulted because

graphophonemic mapping relations were formed but were partial and not as well secured to pronunciations of words in memory.

In the present study, very few learning trials were needed for spelling aids to facilitate word memory (see Figures 1–3). A clear advantage favoring spellings for learning pronunciations was evident by Trial 2 for second and fifth graders. A clear advantage for definitions was evident by Trial 1 for fifth graders and by Trial 2 for second graders. Share (2004) also observed that as few as one learning trial was needed for third graders to retain orthographic information about words in memory. How might such rapid learning be explained? Stuart and Coltheart (1988) identified one possible contributing factor. They suggested that orthographic knowledge provides learners with a basis for expecting specific connections between spoken and written words. For example, students who know how to segment the word *man* into three phonemes and how to spell each phoneme will expect to see the word written *m-a-n* before they ever see it. According to this hypothesis, spelling expectations may support very rapid acquisition because word learning simply entails confirming what is already known.

One striking finding in the present study was that during the learning trials, it was much easier for students to hear words and recall their meanings than to see pictures of the words and recall their pronunciations. Meanings might have been easier to learn because there were only a limited number, 6 in Experiment 1 and 10 in Experiment 2. In contrast, pronouncing words accurately required remembering novel blends of many phonemes, 18 in Experiment 1 and 61 to 63 in Experiment 2, with some phonemes repeated in different words. The difficulty of recalling and pronouncing the correct blends was evident in students' errors. They sometimes mixed up parts from different words—for example, *tamookie* (*tamarack* plus *koomkie*), *friboscis* (*proboscis* plus *fribble*), and *kerfello* (*kerfuffle* plus *scrivello*). In Experiment 2, involving 32 students, this happened 45 times during Trial 1, 61 to 65 times during Trials 2 to 4, and 42 times during Trial 5.

Dual coding provides a theoretical explanation for the ease of learning word meanings in the present study. In their theory, Sadoski and Paivio (2001) distinguished a verbal memory system and a nonverbal imagery system. In the present study, procedures for teaching word meanings utilized both visual and verbal systems. The words were concrete nouns, their meanings were easily constructed from students' experiences, and they were taught with pictures that facilitated the formation of images in memory. Also, the nouns were embedded in verbal definitions and sentence contexts to clarify their position in students' web of linguistic and world knowledge. Sadoski (2005) and Sadoski and Paivio reviewed studies showing that these factors are especially effective for enhancing memory. Dual coding theory's constructs might be incorporated into Ehri's (1992) amalgamation theory to better explain how semantic identities of words are learned.

### Strengths and Limitations

In the present study, the contribution of spelling aids to vocabulary learning was examined via a controlled laboratory task to exclude the confounding or obscuring effects of other variables. Findings clearly reveal that spelling aids provided a major boost. Although these effects were found with direct, individualized instruction, there is no reason to believe that the same boost from



spellings would not also result when vocabulary words are taught to classrooms of students or when students encounter new words in print on their own, provided that the pronunciations and meanings of the written words are apparent to readers. These possibilities should be investigated.

One strength of the present study is that expressive vocabulary acquisition was studied. It is much harder to teach children to produce new words than to teach them to recognize words' meanings (Senechal, 1997). Present findings indicated that showing students spellings of words strengthens their ability to pronounce and use the words in speech.

Another strength of the present study is that a counterbalanced design with random assignment was used to detect the benefits of spelling aids. Students received vocabulary instruction both with and without spelling aids. Almost every student was found to learn words and definitions better when spellings were present than when they were not, regardless of the particular words they learned in each condition and the order in which the two tasks were completed. Therefore, even though the samples were relatively small, the effects were robust.

Some limitations of the study can be identified. The words taught were concrete nouns, which are easier to learn than other parts of speech (Elley, 1989). The words represented concepts already familiar to children rather than unfamiliar concepts. Also, the words' spellings conformed to the English writing system, making them easier to remember than less regularly spelled words (Ehri, 1997). Children's memory for the words was tested after a delay of 1 day, whereas the goal of vocabulary instruction is long term. The participants were second and fifth graders whose parents provided written consent. Further research is needed to show that findings generalize to other types of words, participants, and testing conditions.

Students were sampled from urban, low-socioeconomic status schools with large minority populations, and their vocabulary levels were well below national norms on the PPVT-III test. Nevertheless, it is likely that findings generalize to other populations of students. The Ehri and Wilce (1979) study was conducted with middle class, primarily Caucasian students, and results showed the same benefit of orthography for recall of pronunciations. Biemiller and Slonim (2001) compared children with smaller vocabularies to those with larger vocabularies in Grades 3 through 6 and found that gains in vocabulary growth were comparable in the two groups.

### Implications for Instruction

The present findings carry several important implications for the teaching of vocabulary in schools. One implication involves teachers' and students' awareness of the contribution of orthography. If teachers are not aware of its importance, they may overlook it in their teaching. For example, when they encounter words that their students are unlikely to know as they are teaching a lesson, they may explain the meanings of the words orally without writing them on the board. When they review harder words before or after they have read a text orally to their class, they may define the words orally but not write them. Teachers who make a point of showing the written forms of words as part of vocabulary instruction do so on the basis of common sense rather than on the basis of empirical evidence, which has been meager prior to this study,

or the recommendations of vocabulary instruction experts, who have largely overlooked the contribution of spellings, as reported earlier in this article.

Also, students may overlook the written forms of unfamiliar words as they read text. They may not make the effort to sound out spellings but rather may skip over them and substitute synonyms. In fact, skipping and guessing are commonly taught strategies. The present study provides clear evidence that these are not the best strategies for vocabulary learning and that students need to see, hear, and say unfamiliar words whose meanings are being learned because this strengthens their memory for spellings, pronunciations, and meanings of the new words.

On the basis of the present findings, several procedures can be recommended that hold the promise of making teachers' vocabulary instruction more effective. When teachers explain the meanings of words, they should also show students their written forms, tell students how to pronounce the words, and help students analyze how spellings map pronunciations, including grapheme-phoneme correspondences as well as larger syllabic and morphographic units. These steps should help students retain spellings, pronunciations, and meanings of new vocabulary words in memory.

Hatch and Brown (1995) recommended helping learners retain clear images of written words in memory to combat confusions among similar sounding words, especially for English language learners. As evidence, they cited numerous examples of mistakes made by students of English as a second language in defining English words—for example, a Spanish speaker who defined *happened* as *felices*, meaning *happy*. If students had seen the spelling and had possessed sufficient knowledge of the orthographic system, they might not have considered this meaning. The value of spellings for learning a second language may be especially high because graphemes help to clarify the particular phonemes and morphemes in words, thereby securing a more precise representation in memory. Many second-language learners report heavy reliance on orthography for building their vocabularies (Hatch & Brown, 1995).

Researchers studying foreign language acquisition have recognized the contribution of orthography to vocabulary learning. Sparks et al. (1997) examined high school students learning a foreign language. They found that word decoding skill was the best predictor of year-end foreign language oral proficiency, better even than students' grades in their foreign language class the previous year. The authors interpreted their findings to show the utility of print for building representations of spoken language in memory.

Another possible implication of the present findings involves the importance of systematic phonics instruction. This approach to teaching beginning reading may provide the foundation for more effective vocabulary learning by developing students' knowledge of the orthographic system. This, in turn, should improve their ability to learn and remember new vocabulary words, as shown in the present study. This speculation invites more research.

A third instructional implication involves the value of teaching a learning strategy to enhance students' expressive vocabulary. Older students are thought to learn the bulk of new words through reading (Chall et. al., 1990), and much of this is conducted silently. The findings of Shu, Anderson, and Zhang (1995) confirm that students learn new word meanings when they read text silently.



However, it is not clear whether they learn the pronunciations of new words as well, for this was not studied. A question of interest given the present findings is whether pronunciations are learned along with meanings when text is read silently and whether instructors could enhance this learning by teaching students to stop and say aloud any unknown words they encounter. According to the present findings, this strategy should enhance grapheme-phoneme processing and, hence, the retention of pronunciations along with meanings in memory. This possibility invites further research.

To conclude, we quote from an anonymous reviewer's evaluation of our study:

The research conducted in this study makes a significant contribution to reading theory and practice in that it ties together the orthographic/phonological system with the meaning system in ways not previously thought. The researchers show quite clearly that students who see the spellings of words actually learn the meanings of the words more easily. This is quite a remarkable finding. I know of no research on vocabulary learning or vocabulary instruction that rests on such a claim. The emphasis in the vocabulary literature is on learning the meanings of words, not on the value of seeing the words and tying together knowledge from the different systems. Because of this connection, the research makes a significant contribution to reading theory and also has specific and practical application to everyday practice in a way that few research studies do.

## References

- Barron, R. (1986). Word recognition in early reading: A review of the direct and indirect access hypotheses. *Cognition*, 24, 93-119.
- Baumann, J., & Kameenui, E. (1991). Research on vocabulary instructions: Ode to Voltaire. In J. Baumann & E. Kameenui (Eds.), *Handbook of research on the teaching of language arts* (pp. 604-632). New York: Macmillan.
- Beck, I., & McKeown, M. (1991). Conditions of vocabulary acquisition. In R. Barr, M. Kamil, P. Mosenthal, & P. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 789-814). New York: Longman.
- Beck, I., & McKeown, M. (1999). Comprehension: The sine qua non of reading. *Teaching & Change*, 6, 197-212.
- Beck, I., & McKeown, M. (2007). Increasing young low-income children's oral vocabulary repertoires through rich and focused instruction. *Elementary School Journal*, 107, 251-271.
- Beck, I., McKeown, M., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. New York: Guilford Press.
- Biemiller, A. (2001). Teaching vocabulary: Early, direct, and sequential. *American Educator*, 25, 24-47.
- Biemiller, A. (2004). Best practices in vocabulary instruction: What effective teachers do. In L. Morrow, L. Gambrell, & M. Pressley (Eds.), *Best practices in literacy instruction* (pp. 87-110). New York: Guilford Press.
- Biemiller, A. (2006). Vocabulary development and instruction: A prerequisite for school learning. In D. Dickinson & S. Neuman (Eds.), *Handbook of early literacy research* (Vol. 2, pp. 41-51). New York: Guilford Press.
- Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology*, 93, 498-520.
- Blachowicz, C., & Fisher, P. (2004). Vocabulary lessons. *Educational Leadership*, 61, 66-69.
- Boder, E., & Jarrico, S. (1982). *Boder Test of Reading and Spelling Patterns*. San Antonio, TX: Harcourt Brace.
- Carey, S. (1978). The child as a word learner. In M. Halle, J. Bresnan, & G. Miller (Eds.), *Linguistic theory and psychological reality* (pp. 264-293). Cambridge, MA: MIT Press.
- Chall, J., Jacobs, V., & Baldwin, L. (1990). *The reading crisis*. Cambridge, MA: Harvard University Press.
- Cohen, J. (1988). *Statistical power analysis for the behavior sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Dunn, L., & Dunn, L. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Service.
- Ehri, L. (1984). How orthography alters spoken language competencies in children learning to read and spell. In J. Downing & R. Valtin (Eds.), *Language awareness and learning to read* (pp. 119-147). New York: Springer Verlag.
- Ehri, L. (1985). Effects of printed language acquisition on speech. In D. Olson, N. Torrance, & A. Hildyard (Eds.), *Literacy, language and learning: The nature and consequences of reading and writing* (pp. 333-367). Cambridge, MA: Cambridge University Press.
- Ehri, L. (1991). Development of the ability to read words. In R. Barr, M. Kamil, P. Mosenthal, & P. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 383-417). New York: Longman.
- Ehri, L. (1992). Reconceptualizing the development of sight word reading and its relationship to recoding. In P. Gough, L. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 107-143). Hillsdale, NJ: Erlbaum.
- Ehri, L. (1993). How English orthography influences phonological knowledge as children learn to read and spell. In R. Scholes (Ed.), *Literacy and language analysis* (pp. 21-43). Hillsdale, NJ: Erlbaum.
- Ehri, L. (1997). Learning to read and learning to spell are one and the same, almost. In C. Perfetti, L. Rieben, & M. Fayol (Eds.), *Learning to spell* (pp. 237-269). Hillsdale, NJ: Erlbaum.
- Ehri, L. (1999). Phases of development in learning to read words. In J. Oakhill & R. Bard (Eds.), *Reading development and the teaching of reading: A psychological perspective* (pp. 79-108). Oxford, England: Blackwell Publishers.
- Ehri, L. (2005). Development of sight word reading: Phases and findings. In M. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 135-154). Oxford, England: Blackwell.
- Ehri, L., & Wilce, L. (1979). The mnemonic value of orthography among beginning readers. *Journal of Educational Psychology*, 71, 26-40.
- Elley, W. (1989). Vocabulary acquisition from listening to stories. *Reading Research Quarterly*, 24, 174-188.
- Ganske, K. (2000). *Word journeys: Assessment-guided phonics, spelling, and vocabulary instruction*. New York: Guilford Press.
- Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, 27, 513-543.
- Gathercole, S., & Baddeley, A. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. *Journal of Memory and Language*, 28, 200-213.
- Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language*. Hove, England: Erlbaum.
- Goodman, K. (1976). Reading: A psycholinguistic guessing game. In H. Singer & R. Ruddell (Eds.), *Theoretical models and processes of reading* (2nd ed., pp. 497-508). Newark, DE: International Reading Association.
- Gough, P. (1972). One second of reading. In J. Kavanagh & I. Mattingly (Eds.), *Language by ear and by eye* (pp. 331-358). Cambridge, MA: MIT Press.
- Gough, P., & Tunmer, W. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, 7, 6-10.
- Harris, A., & Jacobson, M. (1982). *Basic reading vocabularies*. New York: Macmillan.
- Hart, B., & Risley, T. (1995). *Meaningful differences in the everyday experiences of young American children*. Baltimore: Brookes.
- Hatch, E., & Brown, C. (1995). *Vocabulary, semantics, and language education*. Cambridge, MA: Cambridge University Press.
- Nagy, W., & Scott, J. (2000). Vocabulary processes. In M. Kamil, P.

- Mossenthal, P. Pearson, & R. Barr (Eds.), *Handbook of reading research* (Vol. 3, pp. 269–284). Mahwah, NJ: Erlbaum.
- National Reading Panel. (2000). *Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. Rockville, MD: NICHD Clearinghouse.
- Perfetti, C. (1992). The representation problem in reading acquisition. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 145–174). Hillsdale, NJ: Erlbaum.
- Rack, J., Hulme, C., Snowling, M., & Wightman, J. (1994). The role of phonology in young children's learning of sight words: The direct-mapping hypothesis. *Journal of Experimental Child Psychology*, 57, 42–71.
- Rack, J., Snowling, M., & Olson, R. (1992). The nonword reading deficit in developmental dyslexia: A review. *Reading Research Quarterly*, 27, 28–53.
- Sadoski, M. (2005). A dual coding view of vocabulary learning. *Reading and Writing Quarterly*, 21, 221–238.
- Sadoski, M., & Paivio, A. (2001). *Imagery and text: A dual coding theory of reading and writing*. Mahwah, NJ: Erlbaum.
- Senechal, M. (1997). The differential effect of storybook reading on preschoolers' acquisition of expressive and receptive vocabulary. *Journal of Child Language*, 24, 123–138.
- Share, D. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, 55, 151–218.
- Share, D. (1999). Phonological recoding and orthographic learning: A direct test of the self-teaching hypothesis. *Journal of Experimental Child Psychology*, 72, 95–129.
- Share, D. (2004). Orthographic learning at a glance: On the time course and developmental onset of self-teaching. *Journal of Experimental Child Psychology*, 87, 267–298.
- Shu, H., Anderson, R., & Zhang, H. (1995). Incidental learning of word meanings while reading: A Chinese and American cross-cultural study. *Reading Research Quarterly*, 30, 76–95.
- Snow, C., Burns, M., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academic Press.
- Sparks, R., Ganschow, L., Patton, J., Artzer, M., Siebenhar, D., & Plageman, M. (1997). Prediction of foreign language proficiency. *Journal of Educational Psychology*, 89, 549–561.
- Stanovich, K. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, 21, 360–407.
- Stuart, M., & Coltheart, M. (1988). Does reading develop in a sequence of stages? *Cognition*, 30, 139–181.
- Swanborn, M., & deGlopper, K. (1999). Incidental word learning while reading: A meta-analysis. *Review of Educational Research*, 69, 261–285.
- Thorndike, E., & Lorge, I. (1972). *The teacher's word book of 30,000 words*. New York: Teachers College Press.
- Torgesen, J., Wagner, R., & Rashotte, C. (1999). *Test of Word Reading Efficiency (TOWRE)*. Austin, TX: PRO-ED Inc.
- Trelease, J. (2006). *The read aloud handbook*. New York: Penguin Books.
- Woodcock, R. (1987). *Woodcock Reading Mastery Tests—Revised*. Circle Pines, MN: American Guidance Service.

Received April 13, 2006

Revision received July 10, 2007

Accepted July 26, 2007 ■



# Early Identification of Reading Difficulties Using Heterogeneous Developmental Trajectories

Christy Kim Boscardin and Bengt Muthén  
University of California, Los Angeles

David J. Francis  
University of Houston

Eva L. Baker  
University of California, Los Angeles

Serious conceptual and procedural problems associated with current diagnostic methods call for alternative approaches to assessing and diagnosing students with reading problems. This study presents a new analytic model to improve the classification and prediction of children's reading development. Growth mixture modeling was used to identify the presence of 10 different heterogeneous developmental patterns. In all, 411 children in kindergarten through Grade 2 from 3 elementary schools in Texas were administered measures of phonological awareness, word recognition, and rapid naming skills 4 times a year. The mean ages were 5.8 years ( $SD = 0.35$ ) for the kindergartners, 6.9 years ( $SD = 0.39$ ) for Grade 1, and 8.0 years ( $SD = 0.43$ ) for Grade 2; the percentage of boys was 50%. The results indicate that precursor reading skills such as phonological awareness and rapid naming are highly predictive of word reading (word recognition) and that developmental profiles formed in kindergarten are directly associated with development in Grades 1 and 2. Students identified as having reading-related difficulties in kindergarten exhibited slower development of word recognition skills in subsequent years of the study.

*Keywords:* reading development, screening, reading skills, achievement, longitudinal studies

Recently, with mounting evidence for early detection as the key to remediation and prevention for later reading difficulties, there has been a growing interest and urgent call for earlier identification of children with reading difficulties. According to research funded by the National Institute of Child Health and Human Development (National Reading Panel, 2000), if intervention is delayed until 9 years of age (the time when most children with reading difficulties typically receive services), approximately 75% of children will continue to have reading difficulties in later grades (Lyon, 1998). The majority of children identified as having reading difficulties in early grades will continue to have problems in later grades without appropriate instructional intervention (Juel, 1988; Scarborough, 1998). Satz and Fletcher (1988) suggested that interventions are most effective if implemented prior to the overt manifestation of disability. Schenck, Fitzimmons, Bullard, Taylor, and Satz (1980) also concluded that children at high risk who received intervention

early demonstrated significant improvement in academic performance over time. Earlier studies have shown that older children who were identified as having reading difficulties would not have required learning disability status if their difficulties had been diagnosed and they had received intervention at an early age (De Hirsh, Jansky, & Langford, 1966; Strag, 1972).

Previous studies have shown that overall academic success in later grades can be predicted with reasonable accuracy by using reading outcomes at early grades (Slavin, 1994; Strag, 1972; Torgesen & Wagner, 2002). Previous longitudinal studies have also suggested that children at risk for reading difficulties can be identified much earlier than previously thought (Juel, 1988; Shaywitz, Escobar, Shaywitz, Fletcher, & Makuch, 1992).

## Identification With IQ Discrepancy-Based Method

Despite the importance of early detection, previous methods for identifying children with reading difficulties suffered from the lack of a theoretical foundation and supportive evidence for validity, which unnecessarily delayed identification (Lyon et al., 2001). Previously, children were identified as having reading difficulties if there was a substantial discrepancy between a child's aptitude, typically operationalized by IQ, and his or her reading performance (Francis, Fletcher, Shaywitz, Shaywitz, & Rourke, 1996; Gunning, 1998). Although the IQ discrepancy-based method has been the most widely used definition of reading difficulty, there were several conceptual and measurement problems that warranted an alternative method of identification of persons with dyslexia and other poor readers (Francis, Shaywitz, Stuebing, Shaywitz, &

---

Christy Kim Boscardin and Eva L. Baker, Graduate School of Education and Information Studies and Center for Research and Evaluation, Standards and Student Testing, University of California, Los Angeles; Bengt Muthén, Graduate School of Education and Information Studies and Department of Statistics, University of California, Los Angeles; David J. Francis, Department of Psychology and the Texas Institute for Measurement, Evaluation, and Statistics, University of Houston.

Correspondence concerning this article should be addressed to Christy Kim Boscardin, University of California, Los Angeles Graduate School of Education and Information Studies, 300 Charles E. Young Drive North, Room 101C, 319 GSE&IS/Mailbox 951522, Los Angeles, CA 90095-1522. E-mail: christyk@ucla.edu

Fletcher, 1996; Shaywitz et al., 1992). Dissatisfaction with previous approaches to identification of children with reading difficulties led to consideration of alternative approaches such as Response to Intervention (RTI) in the recent reauthorization of the Individuals with Disabilities Education Act. The core concept of RTI is based on the premise that a student exhibiting a slower rate of development and failure to respond adequately to intervention may be identified as requiring special services and being at risk for learning disability (Fletcher, Coulter, Reschly, & Vaughn, 2004). With RTI, the focus is on screening, instructional intervention, and continual monitoring. Proponents of the RTI approach suggest that with continuous progress monitoring, the focus will be shifted to prevention and intervention rather than relying on the current "wait-to-fail" model facilitated by the use of IQ-discrepancy approaches (Compton, Fuchs, Fuchs, & Bryant, 2006). Early and accurate identification of students at risk for reading difficulties will be fundamental to successful implementation of the various RTI models proposed (Bradley, Danielson, & Hallahan, 2002; Compton et al., 2006; Fuchs, Mock, Morgan, & Young, 2003). As illustrated by use of the IQ discrepancy-based method in the past, using test scores from a single time point is often unreliable and insufficient for identification of reading difficulties.

Early identification of children with reading difficulties will require a system that accurately predicts which children are at risk for reading failure. However, previous screening procedures have yielded unreliable results due to high rates of classification and measurement errors (Fletcher & Satz, 1984; Jenkins & O'Connor, 2002; Scarborough, 1998). Speece (2005) suggested that one of the reasons for the rate of inaccuracy and the problems associated with early identification is lack of consideration and disregard for potential change and growth in the reading development process. Considering that growth and development are fundamental to the concept of learning, it seems only logical to consider growth (longitudinal) data as the primary source for the identification of reading problems.

### Predictors of Reading

Much progress has been made in the past 20 years in understanding the correlates of reading development and the key predictors of reading outcomes (National Reading Panel, 2000; National Research Council, 1998). One of the most significant predictors of early difficulties in acquiring accurate and fluent word recognition skills has been identified as the individual differences in phonological skill (Jenkins & O'Connor, 2002; Liberman, Shankweiler, & Liberman, 1989; Parrila, Kirby, & McQuarrie, 2004). There seems to be a wide consensus that deficits in phonological awareness are related to later reading difficulties (Catts & Kamhi, 1999; Stanovich & Siegel, 1994; Torgesen, Wagner, Rashotte, Burgess, & Hecht, 1997). More recently, some studies have suggested that deficits in both phonological awareness and in serial naming speed may produce more severe reading difficulties, providing evidence in favor of the double-deficit model (Morris et al., 1998; Schatschneider, Carlson, Francis, Foorman, & Fletcher, 2002; Wolf & Bowers, 2000). However, investigation of the relationship between these two skills on reading development over time and the relative importance of these two skills in early identification of children with reading difficulties

has been limited. In this study, we examined the relationship between the development of precursor skills such as phonological awareness and rapid naming and the development of word recognition skill in later grades.

In light of the current call for alternative approaches to identification of students with reading difficulties, we offer a novel approach to examining reading development through the application of a new longitudinal clustering technique called *growth mixture modeling*. The purpose of this study was to determine whether distinctive groups of students with various developmental profiles can be identified based on precursor reading skills and whether these profiles will help characterize the course of reading difficulties manifested during different developmental periods. Analysis of individual growth curves based on early reading skills is an alternative statistical approach that will potentially be helpful in earlier identification of students with reading difficulties. These individual growth curves may not only yield earlier diagnosis of reading difficulties but may also provide contextual data for further analysis of individual differences and problems.

### *Growth Modeling*

Conventional growth modeling has been a useful technique for examining individual differences in learning development (Bryk & Raudenbush, 1992; Jennrich & Schluchter, 1986; Laird & Ware, 1982; Lindstrom & Bates, 1988; Muthén, Khoo, Francis, & Boscardin, 2002). To study development or change over time, the researcher represents individual outcomes using rates of growth over multiple time points. In growth curve models, individual reading development can be formulated in terms of initial reading level and the rate of learning or development of reading. Typically, the analysis of individual change is represented using random coefficient modeling. Individual change and growth over time can be represented through growth models in order to provide a more dynamic view of reading development.

### *Growth Mixture Modeling*

More recently, a new modeling technique that takes into consideration the effects of heterogeneity in a sample to provide more reliable predictions for later development has been introduced (Muthén, 2000; Muthén et al., 2002; Muthén & Muthén, 2000). Because the conventional growth modeling approach estimates the variation in growth curves as a function of growth factors, interactions of growth factors are difficult to detect and model. In contrast, in the growth mixture modeling framework, interactions between heterogeneous individual growth curves and background factors can be examined without the restrictions of conventional growth modeling techniques. Because growth mixture models allow for the effect of covariates to differ based on the developmental profiles, researchers can determine the relationship between rapid naming skills or a student's background on his or her developmental trajectory. Consequently, the application of growth mixture modeling is extremely useful for developing intervention protocols, as well as identifying specific problems related to development of particular reading skills and to other factors that may contribute to individual differences in development. Growth mixture modeling offers significant advantages over conventional



growth modeling techniques (Muthén, 2000; Muthén et al., 2002). Muthén et al. provided precursory examination of how growth mixture modeling can be used for the identification of several different developmental profiles. In the current study, we expanded on the previous studies on the utility of growth mixture modeling to determine whether distinctive groups of students with various developmental profiles can be identified. Specifically, given the importance of phonological awareness skills in early readers, we examined (a) the relationship between initial status and development of phonological awareness during early stages of reading development and of word recognition skills as well as (b) the contribution of rapid naming skills in differentiating poor readers.

Method

Sample

The sample for the current study was drawn from a larger study consisting of 945 students from a cohort-sequential longitudinal study designed to assess the development of early reading skills (Schatschneider et al., 2002; Schatschneider, Fletcher, Francis, Carlson, & Foorman, 2004; Schatschneider, Francis, Foorman, & Fletcher, 1999). The cohort-sequential longitudinal design for sample selection is presented in Table 1. The larger study sample represented a random selection of children in kindergarten through Grade 2 from three elementary schools in Texas. The students represented a random selection (80%) of students who had parental consent to participate in the study. Students with evidence of severe emotional problems, uncorrected vision problems, hearing loss, acquired neurological disorders, or classification at the lowest level of English as a second language based on school designation were excluded from the study. Children referred for special services in kindergarten were also excluded from participating in this study. Students with later referrals for special education services, however, were included in the study. Measurements of early reading skills were taken four times a year from kindergarten through Grade 2. During a given academic year, students were tested four times (October, December, February, and April) on measures related to reading skills.

For our research purposes, the 411 students with complete data for at least kindergarten and Grade 1 were selected out of the original 945 students for the present study. There were no statistically significant differences between the original sample and this subset on any of the measures. Out of these 411 students, only 208 students had complete data for all three grades. Loss of Grade 1 and Grade 2 in this subset was due to the termination of the study, which prevented us from following some students into Grade 2, and one school's decision not to participate in Year 4. For the

subset of children who had at least complete data in kindergarten and Grade 1 ( $n = 411$ ), about 50% of the sample were boys. The ethnic breakdown of the subset sample was 55% White, 17% African American, 16% Hispanic, 11% Asian, and 1% other ethnicity. The mean ages were 5.8 years ( $SD = 0.35$ ) for the kindergartners, 6.9 years ( $SD = 0.39$ ) for Grade 1, and 8.0 years ( $SD = 0.43$ ) for Grade 2. For the purposes of the present study, White and Asian students were considered to be nonminority, whereas Black, Hispanic, and other students were categorized as minority. The Hollingshead (1975) Four Factor Index of Social Status survey administered to parents was used to collect data on socioeconomic status (SES): (a) 8% were classified as lower class, (b) 40% as working class, (c) 45% as middle-upper class, and (d) the remaining 7% did not provide these data.

Measures

A skills assessment battery was administered individually to each child in October, December, February, and April between kindergarten and Grade 2. Kindergarten students were assessed in phonological awareness and rapid naming. In first and second grades, students were assessed in word recognition. Descriptive statistics on the measures are shown in Table 2.

Phonological Awareness

The phonological awareness test that was administered was an experimental version of the Comprehensive Test of Phonological Processes developed by Wagner, Torgesen, and Rashotte (1999). A detailed description of the test is provided in Schatschneider, Fletcher, Francis, Carlson, and Foorman (2004). The assessment consisted of seven subtests, including (a) phoneme segmentation, (b) phoneme elision, (c) sound categorization, (d) first sound comparison, (e) blending onset and rime, (f) blending phonemes into words, and (g) blending phonemes into nonwords.

*Phoneme segmentation.* This task required children to listen to words and then tell the interviewer “each sound you hear in the word in the order that you hear it.” There were 4 practice items and 15 test items, consisting of one- and two-syllable words with two to five phonemes (e.g., *ate*, *up*, *jump*).

*Phoneme elision.* For the test for phoneme elision, the child was asked to say the word after deleting a specific phoneme (e.g., “Say the word *cup*. Now tell me what word would be left if I said *cup* without the /k/ sound”). There were 4 practice items and 15 test items. All phonemes deleted were consonants. The first 12 items were three-phoneme single-syllable words, for which the deletion was at the end of the word for the first 6 items and the beginning of the word for the other 6 items. The last three items were three- to five-phoneme two-syllable words for which the consonant to be deleted was in the middle (e.g., *ti[g]er*).

*Sound categorization.* This task asked a child to select one of four words that did not share a phoneme with the rest (e.g., “In the set of *fun*, *pin*, *bun*, and *gun*, select the one that doesn't sound like the others”).

*First sound comparison.* In this task, the child was asked to point to the picture of the word that begins with the same sound as the presented word. For example, a booklet with pictures of a rug, a saw, and ash was presented, with a target word of *rake*. The

Table 1  
Cohort Structure for the Larger Study

Cohort	Year 1	Year 2	Year 3	Year 4
1	K ( $n = 183$ )	1 ( $n = 170$ )	2 ( $n = 133$ )	
2		K ( $n = 210$ )	1 ( $n = 158$ )	2 ( $n = 91$ )

Note. K = kindergarten; 1 = Grade 1; 2 = Grade 2.

Table 2  
*Descriptive Statistics for the Measures Used in the Study  
 in IRT Scale*

Variable	<i>M</i>	<i>SD</i>
Phonological awareness at Time 1	-1.22	.59
Phonological awareness at Time 2	-1.01	.65
Phonological awareness at Time 3	-0.79	.74
Phonological awareness at Time 4	-0.60	.82
Word recognition at Time 1	-0.98	.84
Word recognition at Time 2	-0.68	.85
Word recognition at Time 3	-0.40	.90
Word recognition at Time 4	-0.12	.90
Word recognition at Time 5	0.26	.76
Word recognition at Time 6	0.46	.77
Word recognition at Time 7	0.60	.77
Word recognition at Time 8	0.77	.77
Rapid naming at Time 4 (transformed)	-0.26	.74

*Note.* Test scores are item response theory (IRT) scores, standardized to have a mean of 0 and a standard deviation of 1.

correct response in this example is *rug*, which corresponds to the first sound of *rake* (Schatschneider et al., 2004).

**Blending onset and rime.** This task required the child to pronounce a word after the onsets (initial consonants or consonant cluster in a syllable) and rimes (vowels and remaining consonants of the word) had been combined. There were 15 test items, with the number of phonemes in the single-syllable words varying from three to four (e.g., *mouse*, *child*).

**Blending phonemes into words.** This task was identical to blending onset and rime except that in this task the child was asked to blend phonemes rather than onsets and rimes. The child was presented with a string of phonemes at a rate of two per second and asked to repeat them by putting the sounds together. There were 6 practice items and 15 test items (one- and two-syllable words) consisting of two to six phonemes (e.g., *i-f*, *t-oy*, *w-a-sh*, *b-a-m-b-oo*).

**Blending phonemes into nonwords.** This task was identical to blending phonemes into words except that nonwords were used in place of real words, with a parenthetical real word rhyme or near rhyme provided as a pronunciation key (e.g., *i-th [with]*, *y-a-s [gas]*, *th-u-ng [rung]*).

Internal consistency estimates for the subtests ranged from .85 to .95 on all occasions. With respect to concurrent validity, the subtests were correlated with the Lindamood Auditory Conceptualization Test (Lindamood & Lindamood, 1979). Correlations ranged from .41 to .75. For the analysis, we combined the scores on these seven phonological awareness tasks into one latent ability score. Instead of using raw total scores of phonological ability, we used item response theory (IRT) scores based on estimates of each person's latent phonological trait to represent phonological awareness with a mean of 0 and a standard deviation of 1.

### Word Recognition

Skills in word recognition were assessed in Grades 1 and 2 by asking students to read aloud 50 words on 4 × 6-in cards. The Grade 1 and Grade 2 lists each consisted of 50 words, with 16 words in common across the two grades. The 50 words included 36

single-syllable, 11 two-syllable, and 3 three-syllable real words. Words were matched for frequency of occurrence (Carroll, Davies, & Richman, 1971) and spanned first- through third-grade levels of difficulty. The internal consistency estimates exceeded .90 in the present study on all occasions. Concurrent and predictive validity for the word list were high, as evidenced by .80 correlations with the Letter Word and Word Attack subtests of the Woodcock-Johnson Psychoeducational Battery-Revised (Woodcock & Johnson, 1989). For the analysis, we used IRT scores based on estimates of each person's latent trait on word recognition skills as an indicator for word recognition ability. IRT estimates were scaled to have a mean of 0 and a standard deviation of 1.

### Rapid Naming

Denckla and Rudel's (1976) Rapid Automatized Naming Tests for Letters was administered in kindergarten. Rapid automatized naming letters were high-frequency lowercase letters (e.g., *a*, *d*, *o*, *s*, *p*). The stimuli consisted of five letters in a row, repeated 10 times in random sequences. The child was asked to name each letter as quickly as possible. The correct number of responses within 60 s was recorded. Test-retest reliability was .57 for kindergarten (reflecting variability in true change over this age range) and .77 for Grades 1 and 2. Children who did not know all five letters were not administered the test (Schatschneider et al., 2004). For the purposes of the current study, the scores obtained from April data collection were log-transformed and included in the analysis.<sup>1</sup>

Descriptive statistics based on the IRT scale on all of the measures are shown in Table 2. For phonological awareness, scores ranged from 1.22 *SD* below the mean to about 0.60 *SD* below the mean across the four time points in kindergarten. For word recognition, scores ranged from 0.98 *SD* below the mean to 0.77 *SD* above the mean across the eight time points during Grades 1 and 2.

### Analysis

The application of growth modeling with mixture components has been explored by other researchers (Nagin, 1999; Verbeke & LeSaffre, 1996); however, recent work introduced by Muthén (2000) and Muthén and Shedden (1999) has provided a much more flexible framework than previous models. General growth mixture modeling introduced by Muthén and colleagues provides technical advantages over conventional growth models by allowing greater flexibility in model specifications and assumptions. One of the theoretical assumptions of the conventional growth model is that the data come from a single population and that the single-population model accounts for all of the variation in the individual trajectories. As both the data and developmental theory suggest, however, there may be several heterogeneous subgroups within this population that require different sets of model specifications and assumptions. For example, as shown in Figure 1, although

<sup>1</sup> For the rapid naming variable, we combined the speed measure, RNL\_S (number correct/number of seconds) with total correct (RNL\_TR) to create TRNL (rapid naming), where  $TRNL = \log_2(RNL\_TR / RNL\_S + .1)$ .



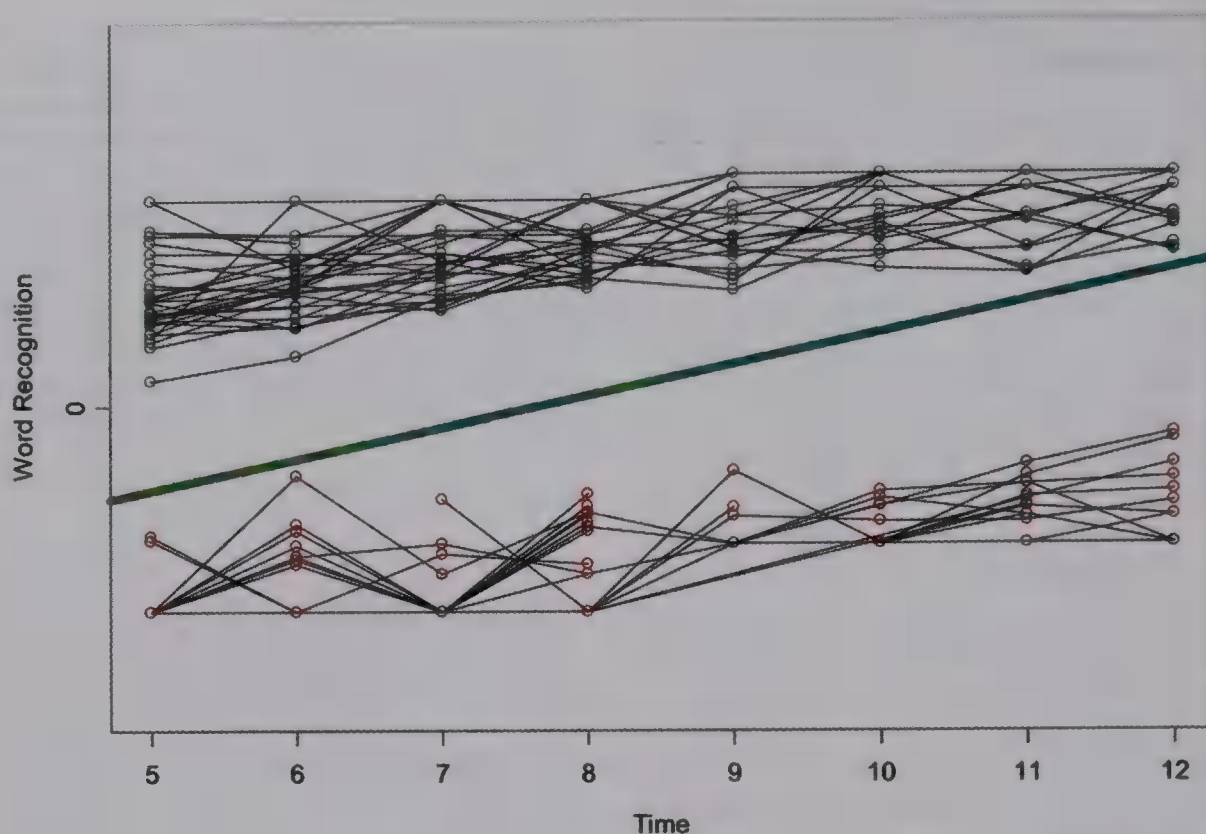


Figure 1. Individual growth trajectories for word recognition development. The two colors represent two distinct growth profiles of reading development.

there may be two different groups of students with very different reading development trajectories, with the conventional growth modeling techniques it would be difficult to detect the misspecification of the model. With the conventional growth models, the estimation of growth is determined by a single collection of growth trajectories with a single vector of means and covariance parameter estimates for intercept and slope parameters. In contrast, growth mixture modeling allows the identification of different subgroups in the model that represent the different collections of reading development trajectories. Hypothetically, the kindergarten growth parameters for a group of students with high intercept (exit level at kindergarten) and growth may influence reading development in Grade 1 differently than for students with low intercept and growth in kindergarten. Accordingly, individuals with similar exit levels in kindergarten may belong to different subgroups with different rates of development. An accelerated growth rate of precursors in kindergarten may suggest that these students are qualitatively different from students with no significant growth during the same time period. Accelerated growth can also be interpreted as a higher aptitude for learning, which could have a greater influence on reading development in Grade 1. Additionally, one could hypothesize heterogeneity in the influence of covariates on the different developmental trajectories.

Growth mixture modeling generalizes conventional growth modeling by allowing heterogeneity of different subgroups in the population through the use of a categorical latent variable (Muthén, 2000, 2001; Muthén & Shedden, 1999). These categorical latent variables or latent classes can represent multiple groups with different developmental trajectories for which group membership is unknown but can be inferred from the data. Individuals are

allowed to be in one of  $K$  latent classes, each with characteristically distinct developmental profiles. Within a class, individual developmental trajectories are allowed to vary around this class profile. For each class  $k$ , continuous outcome variables  $Y$  are assumed to be normally distributed conditional on covariates  $x$ . The growth mixture model can be expressed as follows:

$$Y_{ik} = \nu_k + \Lambda_k \eta_{ik} + K_k x_{ik} + \varepsilon_{ik} \quad (1)$$

and

$$\eta_{ik} = \alpha_k + B_k \eta_{ik} + \Gamma_k x_{ik} + \zeta_{ik}. \quad (2)$$

Here,  $Y$  represents the repeated measures over fixed time points. The  $\eta_{ik}$  are random effects, and  $\Lambda_k$  represents time scores for the shape of the growth curves.  $K_k$  represents the effects of time-varying covariates, and  $\Gamma_k$  represents the effects of time-invariant covariates.  $\alpha_k$  represents the intercepts for  $\eta$  for latent class  $k$ . For the example we discuss,  $\alpha_k$  represents the different reading development trajectories for the different classes. The residual vectors  $\varepsilon_{ik}$  and  $\zeta_{ik}$  are assumed to have covariance matrices  $\Theta_k$  and  $\Psi_k$ , respectively.

The  $K$  trajectory classes are allowed to include variation in both intercepts and slopes in phonological awareness and word recognition. This framework, introduced by Muthén et al. (2002), Muthén and Muthén (2000), and Muthén and Shedden (1999), is much more wide ranging than the mixture models proposed by Nagin (1999), in which it is assumed that  $\Psi_k = 0$  and  $\Theta_k = 0I$ . A model proposed by Muthén (2000) provides more flexibility by allowing class-to-class variation of the covariance matrices  $\Psi_k$  and  $\Theta_k$ . This modeling specification is particularly important when determining

the number of latent class trajectories. Depending on how the degree of class invariance is specified, different values for model fit criteria will be obtained.

Estimated posterior probabilities for each individual's class membership are derived as follows. Define the latent class membership indicators  $c_{ik}$  to be 1 if individual  $i$  belongs to class  $k$ , and 0 otherwise. Then

$$p_{ik} = P(c_{ik} = 1 | y_i, \mathbf{x}_i) \propto P(c_{ik} = 1 | \mathbf{x}_i) f(Y_{ik} | \mathbf{x}_i). \quad (3)$$

In this study, the individual students were assigned to a class based on their highest estimated posterior probabilities. The posterior probabilities were computed for a given individual observation vector  $(y_i, \mathbf{x}_i)$ . In other words, for a given model, individual students' posterior probabilities were computed as a function of the model estimates and the individuals' values on observed variables.

To understand the composition of classes and also to provide stability in class membership, Muthén (2000) introduced a multinomial logistic regression model to represent the relationship between  $c$  (latent class variable) and  $x$  (covariate). The multinomial logistic regression for predicting class membership with a covariate can be expressed as follows:

$$P(c_{ik} = 1 | x_i) = \exp(\beta_{0k} + \beta_{1k}x_i) / \sum_{c=1}^K \exp(\beta_{0c} + \beta_{1c}x_i), \quad (4)$$

for  $k = 1, \dots, K$  where we constrain  $\beta_{0K} = 0$  and  $\beta_{1K} = 0$ , and where  $P(c_{ik} = 1 | x_i)$  is the probability of being in class  $c_{ik}$  conditioned on covariate  $x$ ,  $\beta_{0k}$  is the class intercept, and  $\beta_{1k}$  is the regression coefficient for the  $k$ th class on  $x_i$ , the covariate. The parameter estimates in the model are most easily interpreted by computing predicted class membership probabilities as a function of the covariate.

One of the advantages of this general latent variable modeling framework is that one can systematically explore the influence of precursor skills on later development. For example, students with rapid development of phonological awareness in kindergarten, despite their low performance at the entry level, may continually show rapid development in word recognition as well. In contrast, students with a low entry level and slow development of phonological awareness may not improve much in word reading when they enter Grades 1 and 2. Despite the similarities in initial appearance between these two groups of students, the students with slow development may be qualitatively different from the students with rapid development of phonological awareness. Also, the differences in students' developmental trajectories in kindergarten may differentially influence their later reading development.

### Model Selection and Model Fit

In growth mixture modeling, determination of the optimal number of groups that best represents the data is part of the model selection procedure. The best method for determining the number of classes or groups is still a topic of controversy. For the present study, we considered two statistical indices as well as the overall interpretability of the model based on class counts and substantive theory for model selection. For comparison of nested models with

the same number of classes, the usual likelihood ratio chi-square difference can be used. However, when comparing models with different numbers of classes, the likelihood ratio test may no longer be applicable, and other information criteria must be used. Although McLachlan and Peel (2000) suggested assessing the number of modes of a distribution using the kernel method to estimate the density function, one drawback to this approach is that when classes or the components are not sufficiently separated, the mixture distribution can look unimodal, thus failing to detect the actual number of modes (classes). Also, if the data have a skewed distribution, using a normal mixture model will not be appropriate for capturing the number of classes (McLachlan & Peel, 2000). We used the Bayesian information criterion (BIC) to compare model fit between non-nested models. For a given model, BIC is calculated as follows:

$$\text{BIC} = -2 \log L + r \log n. \quad (5)$$

Here,  $L$  is the value of the model's maximized likelihood,  $n$  is the sample size, and  $r$  is the number of parameters in the model. To determine the optimal number of classes for the best representation of the data, one compares BIC values across the different models, with smaller BIC values indicating a better model fit. However, the overall model selection is guided not only by BIC values but also by entropy indices (described in the next paragraph) and the interpretability of the chosen model, because the BIC tends to favor models with fewer classes (Wiesner & Windle, 2004).

Model selection was also guided by examining the reliability of the classifications via the estimated posterior probabilities of class membership for each individual (Muthén, 2000). The precision of the classification can be assessed by how well the students are being classified into each class. A reliable classification will require the student to have posterior probabilities that are very high for belonging to a single class and very low for belonging to all of the other classes. These probabilities determine the most likely class for each student. For example, a student's estimated probabilities may be .80 for Class 1, .15 for Class 2, .05 for Class 3, and 0 for Classes 4 and 5. A reliable classification is linked to the precision of probabilities in differentiating class membership. To check for precision in classification, one summarizes the probabilities into average posterior probabilities. For example, if the average posterior probability of Class 1 is .90, then one can conclude that the students being classified into Class 1 have, on average, very high probabilities of Class 1 membership. In addition, the quality of the classification is also summarized by the entropy measure (Muthén, 2000). Entropy is expressed as follows:

$$E_K = 1 - \{ \sum_i \sum_k [-p_{ik} \log(p_{ik})] / n \log(K) \}. \quad (6)$$

The expression  $E_K$  is bounded between 0 and 1. An entropy measure close to 1 is considered to be evidence of good classification.

### Analysis Procedure

In the first analysis stage, the development of phonological awareness in kindergarten was examined separately from the development of word recognition, because these represent two distinct (although developmentally sequential) linguistic processes.



The number of classes, as well as the type of growth trajectories, may vary from one grade to the next depending on the different types of latent growth trajectories present. Consequently, a four-group model might be best for representing the developmental profiles in kindergarten, but a five- or six-group model may be more appropriate for representing development in Grades 1 and 2.

Once the number of classes was determined for phonological awareness and word recognition development separately based on the BIC values, the entropy indices, and the interpretability of the model, the final model combining all three grades was selected. The purpose of these two separate analyses in the first stage was to provide a basis for selecting the final model that combined all 3 years of data. Although the developmental trajectories identified in the separate analyses were informative, the overall goal of the study was to find the best representation of data using all three grades. The final model selection was then guided by class counts and the overall model fit.

For the final model, we checked the overall model fit and the quality of the classification by examining how closely the estimates matched the observed data. One way to check this is to compare the estimated mean curve with the observed trajectories of individuals or the observed mean curves based on individual estimated conditional class probabilities. For this technique, we assigned individuals to classes based on the estimated posterior probabilities and then compared the individual trajectories with the estimated mean trajectory (Bandeén-Roche, Miglioretti, Zeger, & Rathouz, as cited in Muthén, 2000). Indication of a good model fit requires close alignment between the individual trajectories and the estimated mean trajectory.

For the present study, we used all 12 time points to investigate reading development across three grade levels (kindergarten, first, and second). For kindergarten, the IRT scale scores on phonological awareness from four different time points (October, December, February, and April) represented reading development. For Grades 1 and 2, we used the IRT scale scores on word recognition spanning 2 years with eight time points to represent continual reading development.

Results

Initial Model Selection Using Two Separate Analyses

Prior to identification of developmental profiles representing all three grade levels, we conducted preliminary analyses for each precursor (phonological awareness) and reading skills (word recognition) separately. First, models with between two and six classes were fit to the longitudinal phonological awareness data from the kindergarten portion of the data. As Table 3 presents, although the four-class model had the lowest BIC value, upon further examination of the class proportions for the four-class model, we determined that the four-class model lacked distinction between the different classes and was difficult to interpret, as shown in Table 4. Given the minimal difference in the entropy values between the four-class and the five-class model, the five-class model was finally chosen for interpretability purposes.

The estimated mean growth curves representing the five different developmental profiles in kindergarten are shown in Figure 2. As Figure 2 illustrates, PA 1 students represented the lowest

Table 3  
*BIC and Entropy Values for Kindergarten Models*

Class	BIC	Entropy
2	1,380.66	0.83
3	1,369.44	0.75
4	1,358.15	0.74
5	1,370.02	0.70
6	1,382.06	0.73

Note. BIC = Bayesian information criterion.

performing group. The estimated means and the standard errors for intercepts and slopes for the five-class model are presented in Table 5. As shown in Table 5, the mean of PA 1 students' phonological scores at the end of kindergarten was about 1.55 SD below the mean, compared to about 1.07 SD below the mean for PA 2 students and 0.26 SD below the mean for PA 3. Both the PA 4 and PA 5 students performed higher, with 0.10 and 1.14 SD above the mean than the overall average for phonological awareness, which was 0.60 SD below the mean at the end of kindergarten. The slope estimate for PA 1 students was 0.03, representing flat trajectory, compared to 0.14 for PA 2, 0.40 for PA 3, 0.27 for PA 4, and 0.34 for PA 5. As shown in Figure 2 and Table 5, PA 1 students started out with low performance in phonological awareness and exhibited no significant improvement throughout the entire kindergarten year. Given the lack of growth during the year, this group of students would be expected to be most at risk for developing reading difficulties in later grades. The estimated means for the slope and intercept for the five-class model are presented in Table 5 and illustrated in Figure 3.

Next, a separate analysis was conducted to fit the word recognition data from Grades 1 and 2. As shown in Table 6, after evaluating different numbers of models using BIC and entropy indices, we found the five-class model to be the best in terms of fit and interpretability. There was a significant decrease in the BIC values between the four-class and the five-class models, however BIC values started to level off after the five-class model. Also, the difference in entropy was minimal across five-class, six-class, and seven-class models. The estimated mean growth curves for the five-class model are shown in Figure 3. After evaluating the BIC values and entropy indices, we determined that the five-class model was the best in terms of fit and interpretability.

As shown in Table 7 and Figure 3, WR 1 had the lowest intercept and slope in the five-class model. The average score for word recognition at the end of Grade 2 was 0.77 SD above the mean. For WR 1 students, the average score was 0.95 SD below the mean, which was significantly lower than the next low performing group (WR 2), which had an average score of 0.46 SD above the mean. Students in WR 1 were thus characterized as the students who were at risk for reading difficulties.

Final Model: Combining Models of Phonological Awareness and Word Recognition

The selected five-class model for phonological awareness and the selected five-class model for word recognition were next combined to allow growth modeling for all 3 years. In the com-

Table 4  
*Class Counts and Proportion of Students in the Kindergarten Four-Class Model*

Variable	Class 1	Class 2	Class 3	Class 4
Class count	229	20	131	31
Proportion of total sample	56	5	32	7

combined analysis, as shown by the path diagram in Figure 4, development of phonological awareness in kindergarten was represented by Intercept 1 and Growth 1, development of word recognition was represented by Intercept 2 and Growth 2, classes were represented as latent variables C1 and C2, and rapid naming measured at the end of kindergarten was also added in the model as a covariate for class membership. Given previous research on phonological awareness and rapid naming, we believed that although there is significant overlap between these two skills, they contribute independently to reading and that rapid naming can be considered an etiologically distinct source of variance in reading outcomes (Petrill, Deater-Deckard, Thompson, DeThorne, & Schatschneider, 2006a).

As an initial step in the analysis, we considered the model in which all students stayed in the same developmental trajectory throughout all 3 years, as shown in Table 8. However, in the final model, to provide a more realistic representation of the data, students were allowed to change class membership during the transition from kindergarten (phonological awareness) to Grades 1 and 2 (word recognition). For example, a student classified as PA 1 based on phonological awareness development was allowed to

Table 5  
*Intercept and Slope for Five Phonological Awareness Development Profiles*

Profile	Intercept <i>M</i> ( <i>SE</i> )	Slope <i>M</i> ( <i>SE</i> )
PA 1	-1.55 (0.26)	0.03 (0.04)
PA 2	-1.07 (0.23)	0.14 (0.06)
PA 3	-0.26 (0.19)	0.40 (0.16)
PA 4	0.10 (0.15)	0.27 (0.04)
PA 5	1.14 (0.31)	0.34 (0.09)

Note. PA = phonological awareness.

progress to a WR 2 or to a WR 3 profile based on word recognition development. By allowing change in class membership, we were able to determine which students stayed in the same developmental trajectories throughout 3 years and which students changed their class membership. The assumption was that if students did stay within the same class throughout the three grades, then the kindergarten classification should correspond directly to Grade 1 and Grade 2 classification. To check for this assumption, we fitted the observed growth curves in Grades 1 and 2 to the estimated mean curves from the kindergarten classes. As illustrated in Figure 5, the observed growth curves in Grades 1 and 2 did not fit the estimated mean curve represented by the dark solid line. Subsequently, we determined that not all PA 1 students corresponded directly to WR 1 students, and not all PA 2 students corresponded directly to WR 2 students, and so on. Instead, as expected, some students did move into other developmental trajectories in Grades 1 and 2. To

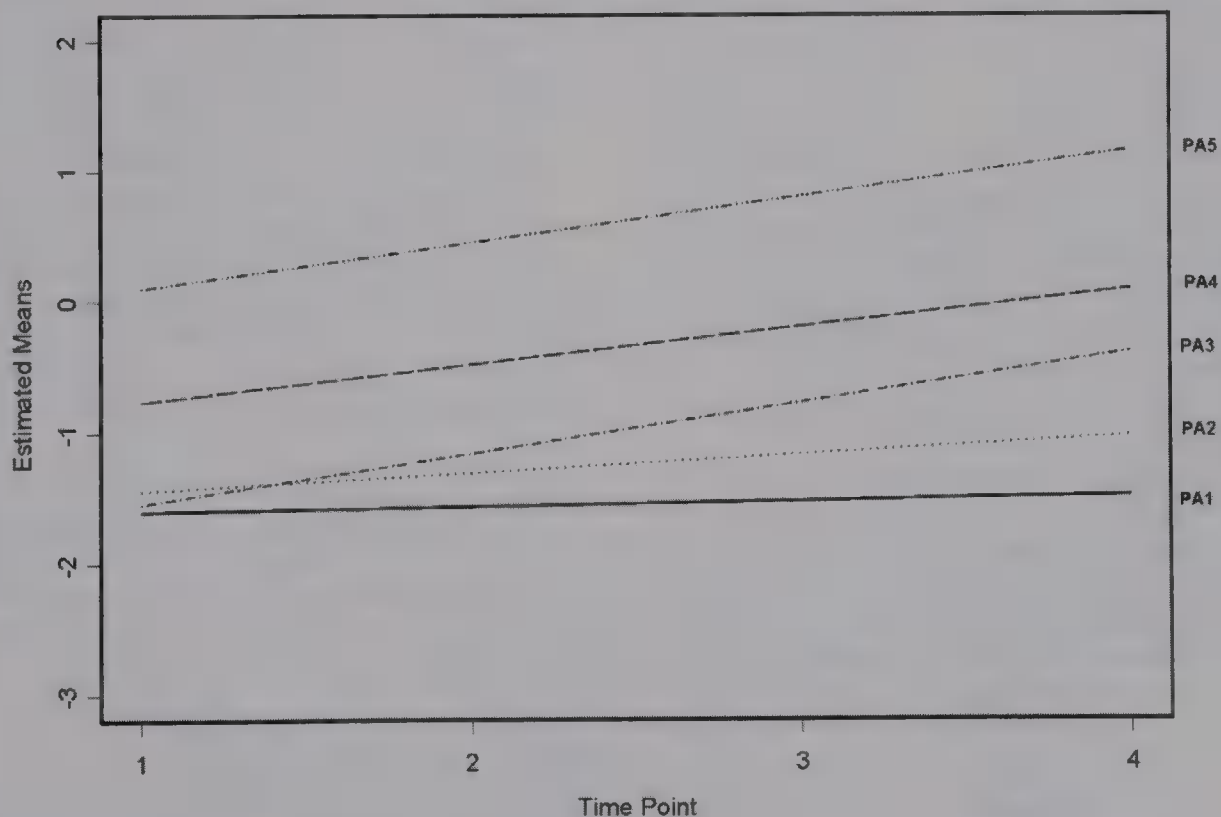


Figure 2. Estimated mean growth curves for the five-class model representing the phonological awareness (PA) in kindergarten.



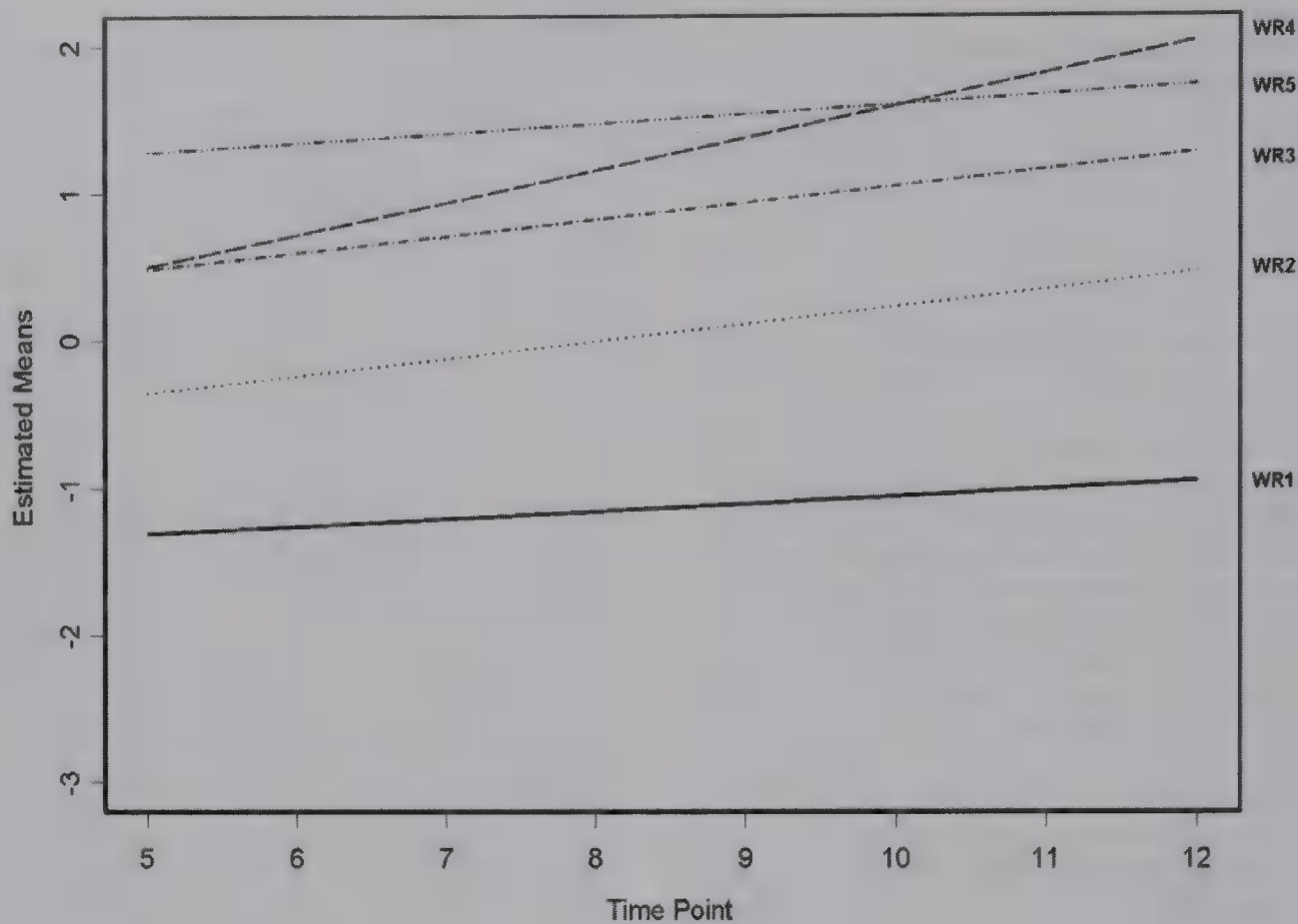


Figure 3. Estimated mean growth curves for the five-class model representing the word recognition (WR) development in first and second grades.

represent this transition from kindergarten to Grades 1 and 2, we allowed for transitional classes in the final model specification.

On the basis of class count, entropy, and model fit, the resulting 10-class model included several different transitional patterns in class membership from kindergarten to Grades 1 and 2. Table 9 describes the 10 different developmental profiles identified in the final model and the class counts based on their estimated class probabilities. The parameter estimates and their standard errors are shown in Table 10 and Figure 6 illustrates the estimated mean curves for phonological awareness and word recognition development. For example, in Figure 6, Class 1 represents the students with developmental profiles corresponding to PA 1 and WR 1, and Class 3 represents the students with developmental profiles corresponding to PA 2 and WR 3.

Table 6  
*BIC and Entropy Values for Grade 1 and 2 Models*

Model	BIC	Entropy
Two-class	2,260.08	0.76
Three-class	2,297.29	0.74
Four-class	2,273.16	0.73
Five-class	2,245.80	0.78
Six-class	2,236.49	0.79
Seven-class	2,233.74	0.82

Note. BIC = Bayesian information criterion.

As illustrated by Figure 6, in the final 10-class model, Class 1 (PA 1 transition into WR 1) students again represented the students who were at risk for reading difficulties because these students' trajectories remained flat. The Growth 1 estimates, which represented the slope for phonological awareness in the final model, ranged from 0.05 (representing flat trajectory) to 0.33 (significant growth). Similarly, for Growth 2, the estimates ranged from 0.03 to 0.27. As shown in Table 10, Class 1 students' reading development trajectories remained flat across all 3 years at 0.05 and 0.03 for Growth 1 and Growth 2, respectively. Class 2 students who, based on the slope and intercept estimates, started out initially at about the same level on phonological awareness skills as Class 1 students, continued to show growth on both phonological

Table 7  
*Intercept and Slope for Five Word Recognition Development Profiles*

Profile	Intercept <i>M (SE)</i>	Slope <i>M (SE)</i>
WR 1	-0.95 (0.10)	0.12 (0.02)
WR 2	0.46 (0.08)	0.27 (0.01)
WR 3	1.27 (0.08)	0.26 (0.01)
WR 4	2.03 (0.14)	0.51 (0.02)
WR 5	1.73 (0.07)	0.15 (0.01)

Note. WR = word recognition.

awareness (Growth 1 = 0.10) as well as word recognition (Growth 2 = 0.23) in contrast to Class 1 students, who showed essentially no growth (Growth 1 = 0.05 and Growth 2 = 0.03) during the same time period. As shown in Table 9, the groups on the diagonal represent the students who remained within the same class across the three grades. The students transitioning into different classes from their original classification in kindergarten are represented by Class 3 (PA 2 into WR 3), Class 4 (PA 3 into WR 2), Class 6 (PA 3 into WR 4), Class 7 (PA 4 into WR 3), and Class 9 (PA 4 into WR 5). The proportion of students that transitioned from one developmental profile to another is shown by the italicized entries in Table 9.

As Table 9 shows, there were only 8 students grouped into Class 4. This class comprised students who were originally classified as PA 3 in kindergarten and who subsequently moved into WR 2 in Grades 1 and 2, which were only a few students. The results also indicated that none of the students transitioned into or out of Class 1. This finding suggests that students in Class 1 were very homogeneous and were indeed the students who we considered to be most at risk for reading difficulties in later grades.

We also investigated the relationship between the growth parameters in kindergarten and the growth parameters in Grades 1 and 2. Only the relationship between word recognition growth and phonological awareness intercept was shown to be statistically significant. In other words, the rate of development in Grades 1 and 2 was directly related to the status of phonological awareness at the end of kindergarten.

#### Evaluation of Model Fit for the 10-Class Model

In order to determine how well this 10-class model fit the data, we plotted a random sample of the observed individual trajectories against the model-estimated mean trajectories. Each individual student was assigned to his or her respective class based on the student's weighted individual class probabilities. A random sample of individual trajectories (observed, not estimated) was plotted

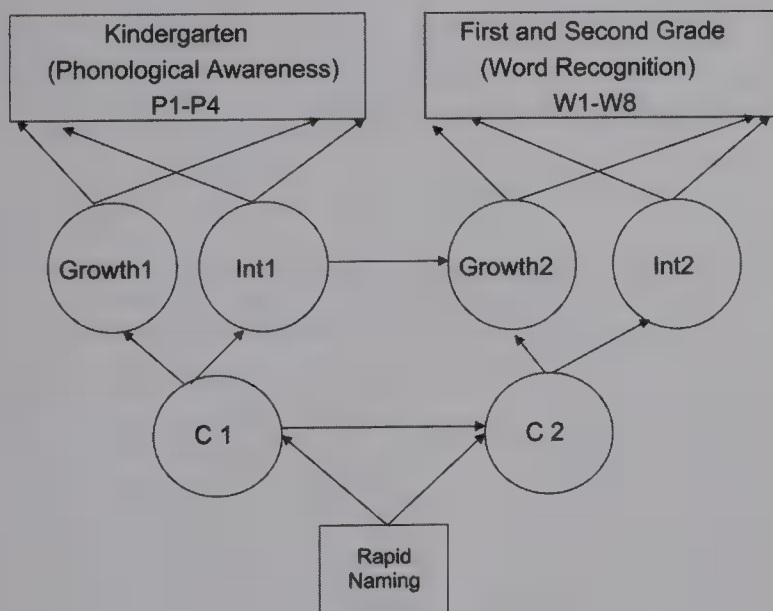


Figure 4. Path diagram for phonological awareness (P) and word recognition (W) combined. Int = intercept; C = class.

Table 8

Five-Class Model Specification

Kindergarten	Grades 1 and 2				
	WR 1	WR 2	WR 3	WR 4	WR 5
PA 1	Class 1	Class 2	Class 3	Class 4	Class 5
PA 2					
PA 3					
PA 4					
PA 5					

Note. WR = word recognition; PA = phonological awareness.

against the estimated mean trajectories for each class for comparison. Except for signs of minor discrepancy in Class 5, all of the observed individual trajectories looked homogeneous around the estimated mean curves (see Figure 7). Class 5 represented the students who were classified as PA 3 in kindergarten and remained in WR 3 in Grades 1 and 2. Additionally, the entropy for the 10-class model was 0.76, which signified moderate to high clarity in the classification.

#### Differences in Rapid Naming

Performance on rapid naming, measured at the end of kindergarten, was included in the model as a covariate for class membership. As explained in the Method section, the significance of this covariate was examined by including a multinomial logistic regression component in the final model. The logistic regression plot shown in Figure 8 illustrates the differential effect of rapid naming on the probability distribution function of the 10 classes. Figure 8 shows that the probability of belonging to Class 1 was much higher when performance on rapid naming was low. Conversely, as illustrated in Figure 8, as the score on rapid naming increased, the probability of an individual belonging to Class 1 decreased significantly. This finding qualitatively differentiated students who entered school with comparably poor phonological awareness (e.g., PA 1 and PA 2 students) but progressed to very different outcomes. The results indicate that students who exhibited no significant improvement in phonological awareness in kindergarten were also the students with the lowest rapid naming skills at the end of kindergarten.

#### Differences in Ethnicity

To help characterize the Class 1 students, we examined the relationships among gender, ethnicity, SES, and class membership using the chi-square test. There was a statistically significant difference in the proportion of minority students in Class 1 compared to other classes. The proportion of minority students in Class 1 was 51% compared to only 31% in other classes ( $p = .01$ ). SES and gender were not statistically significant ( $ps = 1.0$  and  $.34$ , respectively).

#### Discussion

The purpose of this study was to introduce growth mixture modeling as a new approach to the identification of heterogeneous



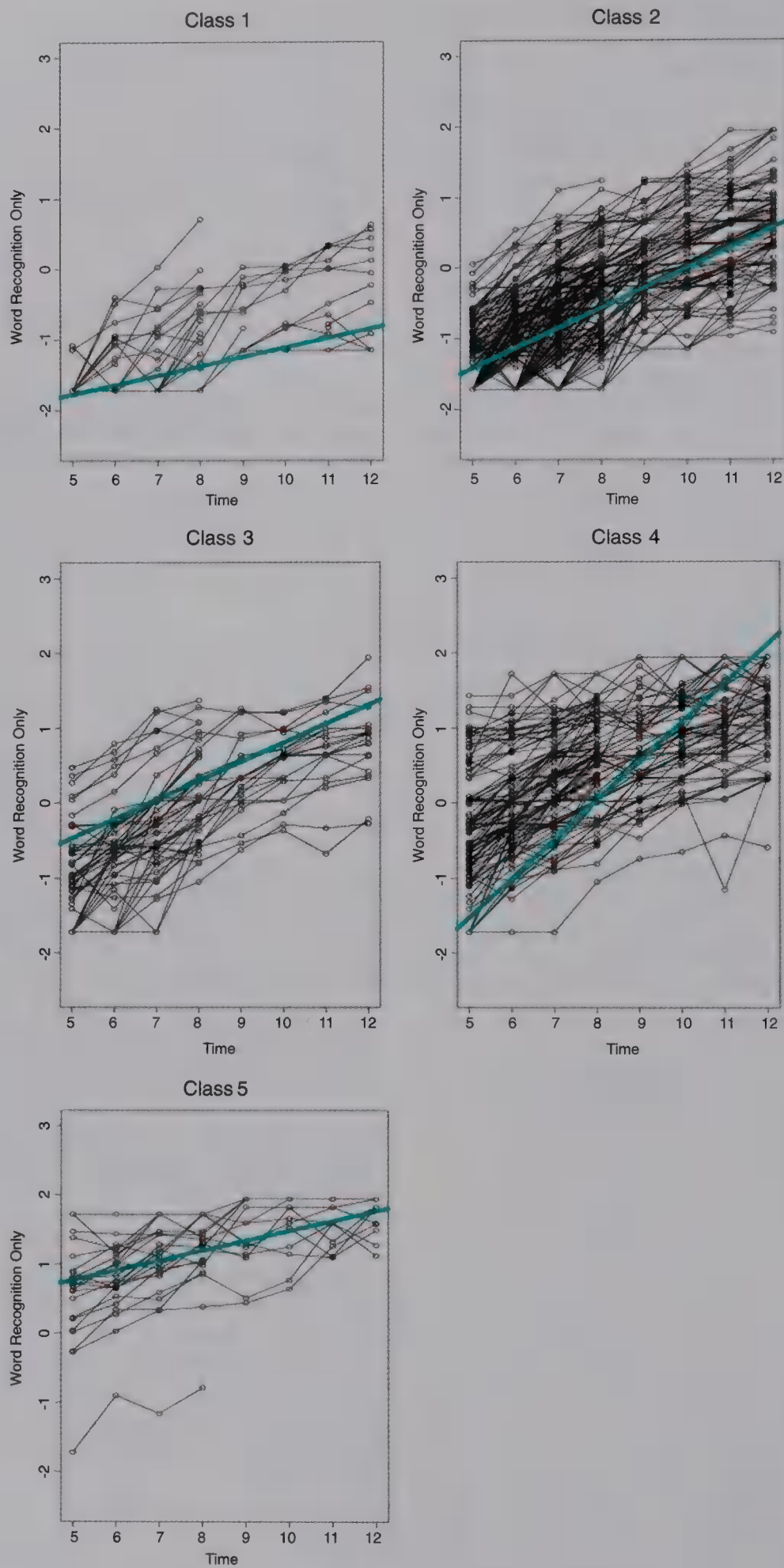


Figure 5. Word recognition development based on kindergarten classification.

Table 9  
*Class Specifications and Class Counts for the 10-Class Model*

Kindergarten	Grades 1 and 2				
	WR 1	WR 2	WR 3	WR 4	WR 5
PA 1	Class 1: 45 (11%)				
PA 2		Class 2: 63 (15%)	Class 3: 77 (19%)		
PA 3		Class 4: 8 (2%)	Class 5: 56 (14%)	Class 6: 36 (9%)	
PA 4			Class 7: 20 (5%)	Class 8: 8 (14%)	Class 9: 30 (7%)
PA 5					Class 10: 18 (4%)

*Note.* Values represent *n* and percentage for each class. The italicized entries represent the five classes that changed class membership between grades. WR = word recognition; PA = phonological awareness.

reading development profiles with data from a 3-year longitudinal study of reading precursors (e.g., phonological awareness) and outcomes (e.g., word recognition). Using this technique, we have identified a group of students with a distinct developmental pattern who are most at risk for reading difficulties. Although further studies are required to validate this group of students as potentially reading disabled, we have empirically identified a group of students with reading difficulties using a new approach.

Application of growth mixture modeling in this study highlights two important issues related to reading development research. First, as shown in previous research, this study empirically demonstrated the multidimensional continuity of the distribution of reading ability (Shaywitz et al., 1992). The results from this study indicate that developmental profiles identified in kindergarten are directly associated with reading development in Grades 1 and 2. The students identified as having difficulties acquiring phonological awareness skills in kindergarten exhibited slower developmental patterns in word recognition skills in subsequent years of the study. Specifically, although students in the lowest performing trajectory class were allowed to change membership with potential for improvement, in fact, nearly all of the students identified as the lowest performing group in kindergarten stayed in the same developmental trajectory throughout the 3 years. The use of growth mixture models to identify and classify students with reading difficulties minimizes anomalies and unfairness that are consequences of using an arbitrary cutoff for classification purposes. Using growth mixture models, researchers can circumvent the

problems associated with arbitrary classification of students as reading disabled.

Previous research has shown (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1994), and the results from this study support the notion, that reading difficulties are best characterized by deficits in prerequisite skills that lead to deficits in reading development, rather than by a lag in reading development. Identification of a group of students with persistent deficits over the 3-year period suggests that unless the students acquire the necessary prerequisite skills, they will continue to lag behind. This finding underscores the need for early identification and interventions specifically targeting deficit skills. Reconceptualizing the identification of reading difficulties using longitudinal measures stimulates further questions regarding the implications of early assessment practices and suggests possible directions for future research in this area. Although there has been considerable debate surrounding the validity of using the Dynamic Indicators of Basic Early Literacy Skills to monitor reading progress, more than 40 states in Reading First schools are now currently using the Dynamic Indicators of Basic Early Literacy Skills to screen Kindergarten–Grade 3 students for potential reading difficulties. Some states are also using the Phonological Awareness Literacy Screening Tests as an alternative tool. Given the lack of consensus on the most appropriate measures for monitoring reading progress, implementation of the proposed identification model will require further research in the area of assessment development. In addition, as researchers consider the practicality of implementing the reading progress monitoring model, the minimum data requirement for reliable classification must also be taken into consideration. As previous research has shown (Rogosa, 1988; Willett, 1988), it becomes difficult to obtain reliable estimates of the correlation between change and initial status with only two assessment data points due to measurement error in initial status.

Second, the findings suggest that the students with reading difficulties may in fact consist of various subgroups or subtypes, each with distinct developmental profiles manifesting from differences not only in outcomes but also possibly in etiology. As shown in the development of phonological awareness in kindergarten, although Class 1 and Class 3 students looked similar in terms of their initial status, rates of development differed significantly between these two groups of students, and, ultimately, this difference in the rate of development manifested a greater gap in students' reading development. Given the significant relationship

Table 10  
*Estimated Means and (Standard Errors) of Intercepts and Growth Parameters for the 10-Class Model*

Class	Intercept 1	Growth 1	Intercept 2	Growth 2
1	−1.50 (0.07)	0.05 (0.02)	−1.05 (0.09)	0.03 (0.04)
2	−1.15 (0.06)	0.10 (0.02)	0.25 (0.14)	0.23 (0.03)
3	−1.15 (0.06)	0.10 (0.02)	1.02 (0.09)	0.27 (0.02)
4	−0.60 (0.22)	0.29 (0.05)	0.25 (0.14)	0.23 (0.03)
5	−0.60 (0.22)	0.29 (0.05)	1.02 (0.09)	0.27 (0.02)
6	−0.60 (0.22)	0.29 (0.05)	1.52 (0.08)	0.26 (0.01)
7	0.15 (0.07)	0.30 (0.02)	1.02 (0.09)	0.27 (0.02)
8	0.15 (0.07)	0.30 (0.02)	1.52 (0.08)	0.26 (0.01)
9	0.15 (0.07)	0.30 (0.02)	1.82 (0.05)	0.17 (0.02)
10	1.09 (0.20)	0.33 (0.05)	1.82 (0.05)	0.17 (0.02)



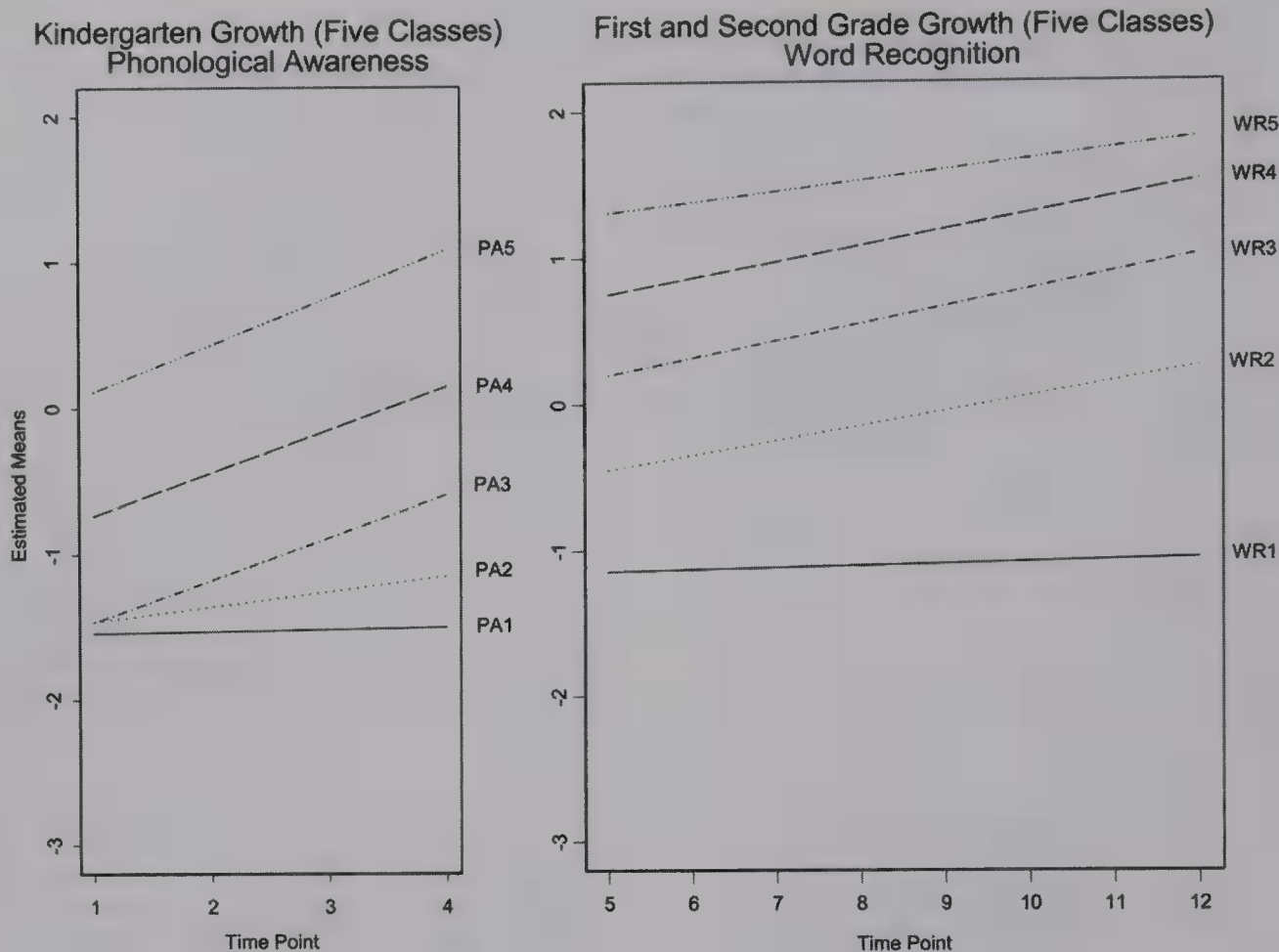


Figure 6. The 10-class model: estimated mean growth curves for phonological awareness (PA) and word recognition (WR).

between class membership and rapid naming, the results suggest that rapid naming and phonological skills are good predictors of subsequent reading development, as was previously shown in other studies (Petrill, Deater-Deckard, Thompson, DeThorne, & Schatschneider, 2006b; Torgesen, Wagner, Simmons, & Laughon, 1990; Walsh, Price, & Gillingham, 1988). Identification of students with poor developmental profiles in reading is a multivariate problem, and a univariate approach is inadequate for studying the complexity of the problem (Fletcher, Francis, Rourke, Shaywitz, & Shaywitz, 1992; Francis et al., 2005). Performance on the rapid naming test in kindergarten was a key indicator for differentiating the different reading development profiles in this study, but other relevant measures such as expressive vocabulary, verbal memory, and story recall should be explored and studied in detail. Although previous studies on reading disability subtypes have used multivariate clustering techniques to identify subtypes, these have been limited to single time point data (Morris et al., 1998). A growth mixture model may provide an alternative approach to identifying reading disability subtypes.

With different developmental profiles of reading, deficits in specific areas can be easily identified, and appropriate instructional strategies can target specific problem areas. As Berninger and Abbott (1994) stated, "The diagnosis of learning disability often is not tied to well-specified deficits clearly linked to instructional interventions" (pp. 166). Consequently, as a large-scale field study

conducted by Haynes and Jenkins (1986) revealed, reading instructional programs often are not linked to meet the needs of the characteristically different individual students. Identification of distinct developmental profiles may provide ways to differentiate treatment based on different treatment responses. With an effective identification system linked to appropriate remediation strategies, students will be able to receive supplementary instruction that is appropriately targeted for maximum benefit. As previous research has suggested (Fletcher, Francis, Morris, & Lyon, 2005; Torgesen et al., 2001), early identification is key to successful remediation programs. It is also important to recognize that, in accordance with the RTI model, identification is only the first stage in the intervention process. For successful remediation and prevention, early screening and appropriate intervention have to be followed up with progress monitoring.

This study also found that the percentage of minority students in Class 1 was higher than in other classes. This increased representation of minorities in Class 1 reflects the fact that minorities are at increased risk for reading problems. A report from the National Research Council (1998) suggested that "children from poor families and children of African American and Hispanic descent are at much greater risk of poor reading outcomes" (p. 27). Teasing apart the relationship of external factors and reading achievement is complicated by inadequate indicators of SES such as the self-report data that we used in our study. One of the limitations of this

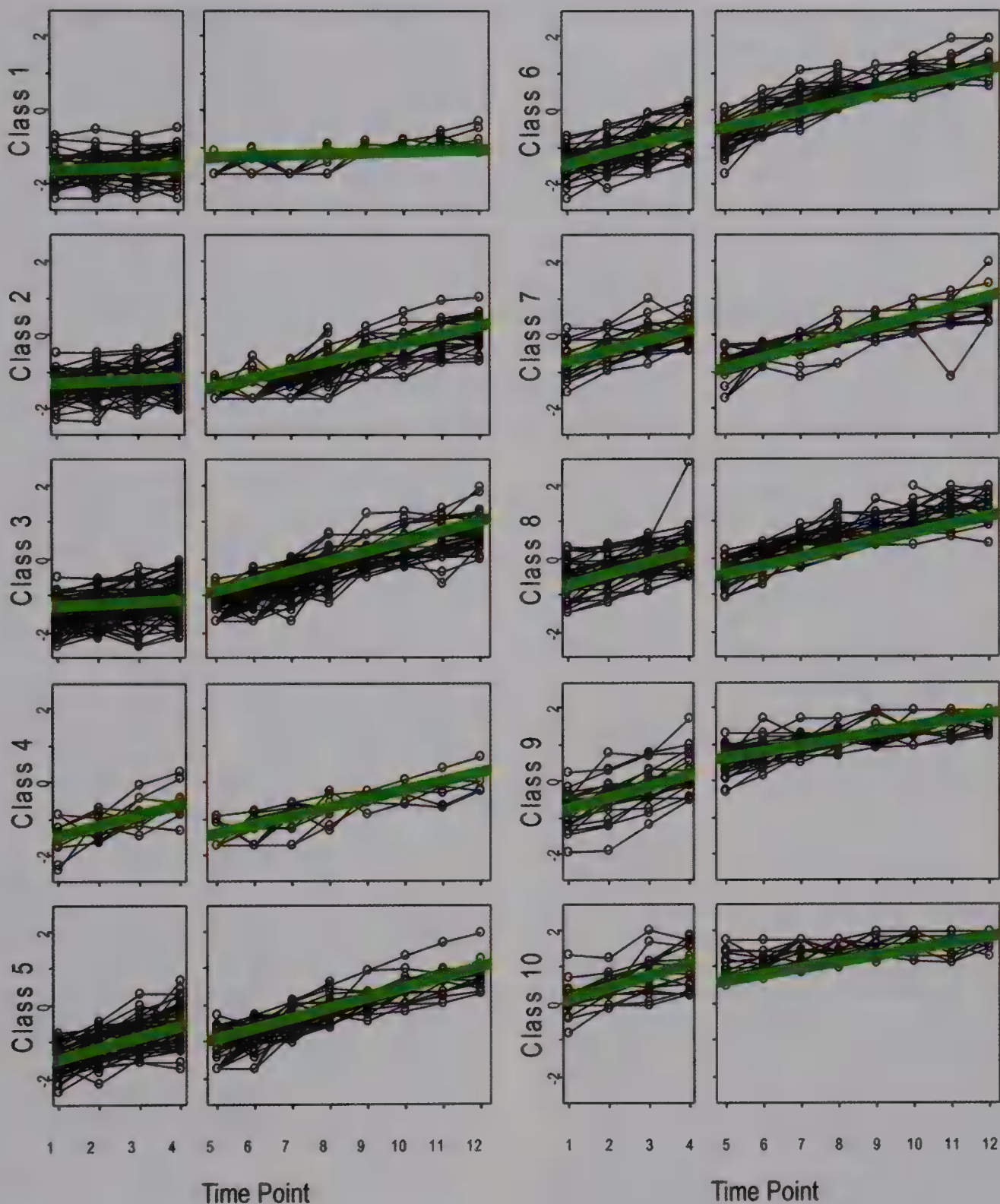


Figure 7. Observed individual growth trajectories with estimated mean growth trajectories for the 10-class model.

study is that, although the identification of Class 1 students was associated with minority status, given the limited information on student background, the characterization of students with potential reading failure was insufficient.

Another limitation of our findings is the relatively restricted measure of word recognition and the lack of a reading comprehension measure. We found that the correlation was quite high between our word recognition measure and the word attack

measure, as measured by the Word Attack subtest of the Woodcock-Johnson Psychoeducational Battery-Revised (Woodcock & Johnson, 1989). At the end of Grade 1, the correlation was .78 between our word reading recognition measure and Woodcock-Johnson Word Attack. At the end of Grade 2, the correlation was .71. Given the importance of reading comprehension skills in reading development, replication of this study with the inclusion of a comprehensive word reading



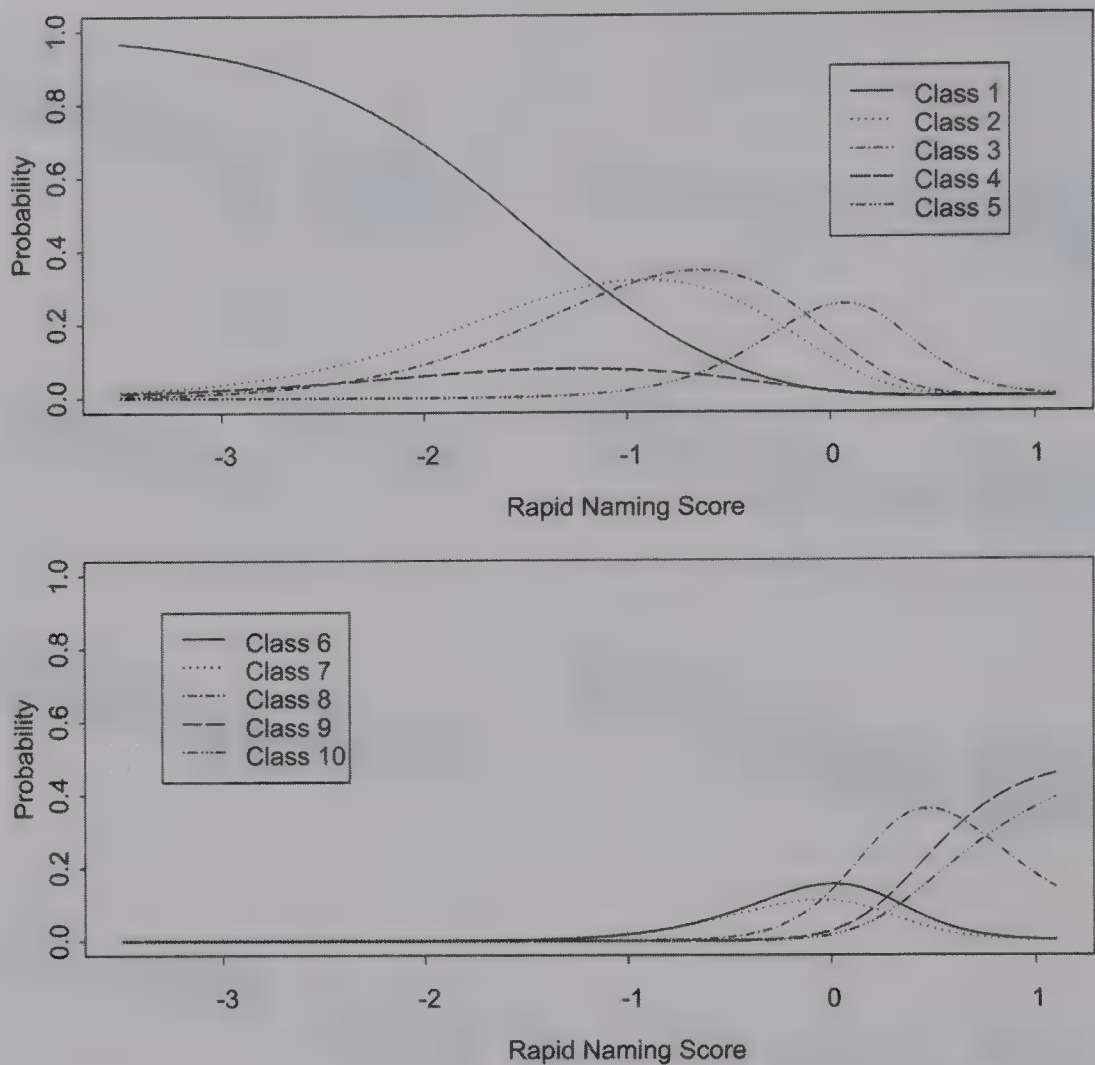


Figure 8. Multinomial logistic regression of class regressed on rapid naming score.

measure that includes nonsense words as well as reading comprehension is warranted.

Another important factor omitted in the analysis was school-level variability. Given there were only three schools in the study, systematic exploration of school effects on student reading outcomes was not possible with this current study sample. The fact that all three schools came from one single district with a common approach to reading instruction should minimize the potential school-level variability due to instruction.

Finally, given the recent findings in behavioral genetics studies (Byrne et al., 2002; Harlaar, Spinath, Dale, & Plomin, 2005; Petrill et al., 2006b), it will be important to consider environmental factors as well as genetic differences in early readers when examining differential treatment effects. Petrill et al. (2006b) found that environmental influences were substantial in explaining the individual variances in phonological awareness, however rapid naming seemed to be significantly influenced by genetic variance. For successful remediation, it is important for future studies to consider various individual and environmental factors to help characterize the different reading development profiles.

References

Berninger, V. W., & Abbott, R. D. (1994). Redefining learning disabilities: Moving beyond aptitude-achievement discrepancies to failure to respond

to validated treatment protocols. In R. G. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues* (pp. 163–185). Baltimore: Brookes.

Bradley, R., Danielson, R.L., & Hallahan, D.P. (2002). *Identification of learning disabilities: Research to practice*. Mahwah, NJ: Erlbaum.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

Byrne, B., Delaland, C., Fielding-Barnsley, R., Quain, P., Samuelsson, S., Høien, T., et al. (2002). Longitudinal twin study of early reading development in three countries: Preliminary results. *Annals of Dyslexia*, 52, 49–74.

Carroll, J., Davies, P., & Richman, B. (1971). *The American heritage word frequency book*. Boston: Houghton Mifflin.

Catts, H. W., & Kamhi, A. G. (1999). *Language and reading disabilities*. Needham Heights, MA: Allyn & Bacon.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology*, 98, 394–409.

De Hirsh, K., Jansky, J., & Langford, W. (1966). *Predicting reading failure*. New York: Harper & Row.

Denckla, M., & Rudel, R. (1976). Rapid automatized naming (RAN): Dyslexia differentiated from other learning disabilities. *Neuropsychologia*, 14, 471–479.

Fletcher, J. M., Coulter, W. A., Reschly, D. J., & Vaughn, S. (2004). Alternative approaches to the definition and identification of learning

- disabilities: Some questions and answers. *Annals of Dyslexia*, 54, 304–331.
- Fletcher, J. M., Francis, D. J., Morris, R. D., & Lyon, G. R. (2005). Evidence-based assessment of learning disabilities in children and adolescents. *Journal of Clinical Child and Adolescent Psychology*, 34, 506–522.
- Fletcher, J. M., Francis, D. J., Rourke, B. P., Shaywitz, S. E., & Shaywitz, B. A. (1992). The validity of discrepancy-based definitions of reading disabilities. *Journal of Learning Disabilities*, 25, 555–561.
- Fletcher, J. M., & Satz, P. (1984). Test-based versus teacher-based predictions of academic achievement: A three-year longitudinal follow-up. *Journal of Pediatric Psychology*, 9, 193–203.
- Francis, D. J., Fletcher, J. M., Shaywitz, B. A., Shaywitz, S. E., & Rourke, B. (1996). Defining learning and language disabilities: Conceptual and psychometric issues with the use of IQ tests. *Language, Speech, and Hearing Services in Schools*, 27, 132–143.
- Francis, D. J., Fletcher, J. M., Stuebing, K. K., Lyon, G. R., Shaywitz, B. A., & Shaywitz, S. E. (2005). Psychometric approaches to the identification of LD: IQ and achievement scores are not enough. *Journal of Learning Disabilities*, 38, 98–108.
- Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1994). The measurement of change: Assessing behavior over time and within a developmental context. In G. R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues* (pp. 29–58). Baltimore: Brookes.
- Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology*, 88, 3–17.
- Fuchs, D., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research and Practice*, 18, 157–171.
- Gunning, T. G. (1998). *Assessing and correcting reading and writing difficulties*. Boston: Allyn & Bacon.
- Harlaar, N., Spinath, F. M., Dale, P. S., & Plomin, R. (2005). Genetic influences on word recognition abilities and disabilities: A study of 7 year old twins. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 46, 373–384.
- Haynes, M., & Jenkins, J. (1986). Reading instruction in special education resource rooms. *American Educational Research Journal*, 23, 161–190.
- Hollingshead, A. B. (1975). *Four-Factor Index of Social Status*. New Haven, CT: Yale University Press.
- Jenkins, J. R., & O'Connor, R. E. (2002). Early identification and intervention for young children with reading/learning disabilities. In R. Bradley, L. Danielson, & D. P. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 99–149). Mahwah, NJ: Erlbaum.
- Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated measures models with structured covariance matrices. *Biometrics*, 42, 805–820.
- Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology*, 78, 243–255.
- Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963–974.
- Liberman, I. Y., Shankweiler, D., & Liberman, A. M. (1989). The alphabetic principle and learning to read. In D. Shankweiler & I. Y. Liberman (Eds.), *Phonology and reading disability: Solving the reading puzzle* (pp. 1–33). Ann Arbor: University of Michigan Press.
- Lindamood, C. H., & Lindamood, P. C. (1979). *Lindamood Auditory Conceptualization Test*. Austin, TX: PRO-ED.
- Lindstrom, M. J., & Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83, 1014–1022.
- Lyon, G. R. (1998). *Overview of reading and literacy initiatives*. Washington, DC: National Institute of Child Health and Human Development.
- Lyon, G. R., Fletcher, J. M., Shaywitz, S. E., Shaywitz, B. A., Torgesen, J. K., & Wood, F. B. (2001). Rethinking learning disabilities. In C. E. Finn, Jr., R. A. J. Rotherham, & C. R. Hokanson, Jr. (Eds.), *Rethinking special education for a new century* (pp. 259–287). Washington, DC: Thomas B. Fordham Foundation and Progressive Policy Institute.
- McLachlan, G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley Interscience.
- Morris, R. D., Steubing, K. K., Fletcher, J. M., Shaywitz, S. E., Lyon, G. R., Shankweiler, D. P., et al. (1998). Subtypes of reading disability: Variability around a phonological core. *Journal of Educational Psychology*, 90, 347–373.
- Muthén, B. (2000). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: New developments and techniques* (pp. 1–33). Mahwah, NJ: Erlbaum.
- Muthén, B. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. In L. M. Collins & A. Sayer (Eds.), *New methods for the analysis of change* (pp. 289–322). Washington, DC: American Psychological Association.
- Muthén, B., Khoo, S. T., Francis, D., & Boscardin, C. (2002). Analysis of reading skills development from kindergarten through first grade: An application of growth mixture modeling to sequential processes. In S. R. Reise & N. Duan (Eds.), *Multilevel modeling: Methodological advances, issues, and applications* (pp. 71–89). Mahwah, NJ: Erlbaum.
- Muthén, B., & Muthén, L. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*, 24, 882–891.
- Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55, 463–469.
- Nagin, D. S. (1999). Analyzing developmental trajectories: A semi-parametric, group based approach. *Psychological Methods*, 4, 139–157.
- National Reading Panel (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institute of Child Health and Human Development.
- National Research Council. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.
- Parrila, R. K., Kirby, J. R., & McQuarrie, L. (2004). Articulation rate, naming speed, verbal short-term memory, and phonological awareness: Longitudinal predictors of early reading development? *Scientific Studies of Reading*, 8, 3–26.
- Petrill, S. A., Deater-Deckard, K., Thompson, L. A., DeThorne, L. S., & Schatschneider, C. (2006a). Genetic and environmental effects of serial naming and phonological awareness on early reading outcomes. *Journal of Educational Psychology*, 98, 112–121.
- Petrill, S. A., Deater-Deckard, K., Thompson, L. A., DeThorne, L. S., & Schatschneider, C. (2006b). Reading skills in early readers: Genetic and shared environmental influences. *Journal of Learning Disabilities*, 39, 48–55.
- Rogosa, D. R. (1988). Myths about longitudinal research. In K. W. Schaie, R. T. Campbell, W. Meredith, & S. C. Rawlings (Eds.), *Methodological issues in aging research* (pp. 171–210). New York: Springer-Verlag.
- Satz, P., & Fletcher, J. M. (1988). Early identification of learning disabled children: An old problem revisited. *Journal of Consulting and Clinical Psychology*, 56, 824–829.
- Scarborough, H. S. (1998). Predicting the future achievement of second graders with reading disabilities: Contributions of phonemic awareness,



- verbal memory, rapid serial naming, and IQ. *Annals of Dyslexia*, 48, 115–136.
- Schatschneider, C., Carlson, C. D., Francis, D. J., Foorman, B. R., & Fletcher, J. M. (2002). Relationship of rapid automatized naming and phonological awareness in early reading development: Implications for the double-deficit hypotheses. *Journal of Learning Disabilities*, 35, 245–256.
- Schatschneider, C., Fletcher, J. M., Francis, D. J., Carlson, C. D., & Foorman, B. R. (2004). Kindergarten prediction of reading skills: A longitudinal comparative analysis. *Journal of Educational Psychology*, 96, 265–282.
- Schatschneider, C., Francis, D. J., Foorman, B. R., & Fletcher, J. M. (1999). The dimensionality of phonological awareness: An application of item response theory. *Journal of Educational Psychology*, 91, 439–449.
- Schenck, B., Fitzimmons, J., Bullard, P. C., Taylor, H. G., & Satz, P. (1980). A prevention model for children at risk for reading failure. In R. M. Knights & D. J. Bakker (Eds.), *Treatment of hyperactive and learning disordered children* (pp. 31–48). Baltimore: University Park Press.
- Shaywitz, S. E., Escobar, M. D., Shaywitz, B. A., Fletcher, J. M., & Makuch, R. (1992). Distribution and temporal stability of dyslexia in an epidemiological sample of 414 children followed longitudinally. *New England Journal of Medicine*, 326, 145–150.
- Slavin, R. E. (1994). *Preventing early school failure: Research, policy and practice*. Needham Heights, MA: Allyn & Bacon.
- Speece, D. L. (2005). Hitting the moving target known as reading development: Some thoughts on screening children for secondary interventions. *Journal of Learning Disabilities*, 38, 487–493.
- Stanovich, K. E., & Siegel, L. S. (1994). Phenotypic performance profile of children with reading disabilities: A regression-based test of the phonological-core variable-difference model. *Journal of Educational Psychology*, 86, 24–53.
- Strag, G. A. (1972). Comparative behavioral ratings of parents with severely mentally retarded, special learning disability, and normal children. *Journal of Learning Disabilities*, 5, 52–56.
- Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K. S., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities*, 34, 33–58.
- Torgesen, J. K., & Wagner, R. K. (2002). Predicting reading ability. *Journal of School Psychology*, 40, 1–26.
- Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Burgess, S., & Hecht, S. (1997). Contributions of phonological awareness and rapid automatic naming ability to the growth of word-reading skills in second to fifth-grade children. *Scientific Studies of Reading*, 1, 161–185.
- Torgesen, J. K., Wagner, R. K., Simmons, K., & Laughon, P. (1990). Identifying phonological coding problems in disabled readers: Naming, counting, or span measures? *Learning Disabilities Quarterly*, 13, 236–243.
- Verbeke, G., & LeSaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random effects population. *Journal of the American Statistical Association*, 91, 217–221.
- Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). *Comprehensive Test of Phonological Processes*. Austin, TX: PRO-ED.
- Walsh, D. J., Price, G. G., & Gillingham, M. G. (1988). The critical but transitory importance of letter naming. *Reading Research Quarterly*, 23, 108–122.
- Wiesner, M., & Windle, M. (2004). Assessing covariates of adolescent delinquency trajectories: A latent growth mixture modeling approach. *Journal of Youth and Adolescence*, 22, 431–442.
- Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education*, 15, 345–422.
- Wolf, M., & Bowers, P. G. (2000). Naming-speed deficits in developmental reading disabilities: An introduction to the special series on the double-deficit hypothesis. *Journal of Learning Disabilities*, 33, 322–333.
- Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-educational Battery-Revised*. Allen, TX: DLM Teaching Resources.

Received February 2, 2006

Revision received June 19, 2007

Accepted July 3, 2007 ■

# The Effects of Tasks on Integrating Information From Multiple Documents

Raquel Cerdán  
Catholic University of Valencia

Eduardo Vidal-Abarca  
University of Valencia

The authors examine 2 issues: (a) how students integrate information from multiple scientific documents to describe and explain a physical phenomenon that represents a subset of the information in the documents; and (b) the role of 2 sorts of tasks to achieve this type of integration, either writing an essay on a question requiring integration across texts or answering shorter intratext questions that require students to integrate information within a single text, while superficial and deep comprehension measurements are obtained. Undergraduate students answered 1 of the 2 types of questions, and their reading times were recorded. Half of the sample thought aloud. Results showed that the integration question increased integration and decreased the processing of isolated units of information, which enhanced deep learning, whereas no differences between the 2 sorts of tasks on memory recall were apparent. This research also provides evidence for the discrepancy between training and posttraining effects (R. A. Schmidt & R. A. Bjork, 1992).

**Keywords:** multiple documents, integration, task effects, thinking aloud, posttraining and deep learning effects

Nowadays, students are frequently exposed to different and multiple sources of information with which they may be asked to perform a variety of tasks such as writing essays, answering comprehension questions, or locating specific units of information. Let us take the example of a science teacher in a high school who gives a group of students a set of reliable documents on bacteria resistance to antibiotics for students to learn some specific issues included in the documents, such as how bacteria resist the effects of antibiotics and which biological mechanisms explain this phenomenon and its transmission to other bacteria. As the documents have been written by different authors with different purposes in mind, the information they contain may partially overlap across texts. At the same time, a given passage will include specific aspects of the problem not considered in a different passage. Additionally, the documents might contain information different from that of the specific interests of the teacher (e.g., the consequences of bacteria resistance or solutions for the problem).

Therefore, the first job for the student is to select and process the pieces of information relevant to the specific issues proposed. He or she also has to recognize when an idea in a document has been presented in a different document, though it may be discussed using different words, so that a reinstatement process occurs. He or she also has to learn that a piece of information from a document complements another piece from a different document and decide what sort of relationship between the two pieces can be estab-

lished. Finally, the reader has to organize all of the selected information to give a coherent answer to the teacher's question. Thus, integrating information from multiple documents is a highly demanding task that involves more processes than those involved when just reading a single text.

Much research on integration of information from multiple documents has been conducted with texts addressing a controversial historical event presented from different and opposing points of view (Britt, Rouet, Georgi, & Perfetti, 1994; Rouet, Britt, Mason, & Perfetti, 1996; Rouet, Favart, Britt, & Perfetti, 1997; Wineburg, 1991, 1994; Wolfe & Goldman, 2005). In this research, models of different situations were formed from the texts, and the students' task was to compare the documents and undertake some sort of reasoning from the models. This process involved constructing a *document model* that consisted of an *intertext model* representing the relationships among the different texts and among a text and elements of the situation (e.g., the source information or rhetorical goals), and a *situations model*, that is, the situations construed from each text and multiple interrelated situations (Perfetti, Rouet, & Britt, 1999).

In this article, in contrast, we are dealing with a different type of integration of information from multiple documents, that is, the description and explanation of a physical phenomenon created by putting together a subset of information from documents. This sort of integration may not need the construction of such a complex document model, a conclusion pointed out by Perfetti et al. (1999). According to these authors, whereas tasks and contexts that orient readers toward the issue of what was said by whom (i.e., comparison among documents) may involve the construction of a complex document model in which intertext and situations models can be distinguished, tasks and contexts that address the issue of what happened or how this could be explained may involve the construction of a single and simpler situation model. Students' success in this task would primarily depend on selecting, processing, and integrating relevant units of information from the documents. This

---

Raquel Cerdán, Department of Educational Psychology, Catholic University of Valencia, Valencia, Spain; Eduardo Vidal-Abarca, Department of Developmental and Educational Psychology, University of Valencia, Valencia, Spain.

This research was conducted with the financial support of the Spanish Ministry of Education under Contract BSO2002-00545.

Correspondence concerning this article should be addressed to Raquel Cerdán, Faculty of Psychology, Catholic University of Valencia, Guillén de Castro, 94, 46003, Valencia, Spain. E-mail: Raquel.Cerdan@ucv.es



simpler integration process has received less attention in the research, and it is this sort of integration that we explore in this article.

The science teacher we are imagining should decide which sort of task is more helpful to students in order to reach their goals. A common approach is to facilitate the student's job as much as possible. However, some authors have demonstrated that conditions that make reading more effortful are more beneficial to learning than are less demanding conditions, though they make no difference to superficial comprehension measures such as text recall. For example, McNamara and colleagues (McNamara, Kintsch, Songer, & Kintsch, 1996; McNamara & Kintsch, 1996) found that less coherent text improved the reader's inferences but not recall. Mannes (Mannes, 1994; Mannes & Kintsch, 1987) found that readers performed better on a problem-solving task when an introductory outline did not match the text well.

Following these ideas, our goal in this research was twofold. First, we were interested in studying the processes of how students integrate information from multiple scientific documents. Second, we wanted to test whether asking students to answer a question that requires them to integrate information across documents is more beneficial in terms of deep learning than dividing the task into smaller intratext questions that may facilitate the student's performance on the task and recall of information but hinder integration across documents. Before getting into the specifics of the research, we would like to highlight important theoretical questions about the two issues mentioned previously, that is, the processes to integrate information from multiple documents and how specific tasks may be more effective than others in fostering such integration.

### Processes to Integrate Information From Multiple Documents

Integrating information seems to be one of the central processes in reading multiple documents, as it helps in constructing a coherent mental representation from different sources. To create coherence in his or her mind, the reader has to undertake some mental activity to establish connections among different units of information from the different texts. In the context of learning from a single text, this connecting process is called *inference making*, and its activation results in the construction of a coherent mental representation (Goldman, 1997; Kintsch, 1998). On-line evidence for integration processes when reading a single text was provided by Coté, Goldman, and Saul (1998), who found that the number of times sentences were accessed was a strong predictor of students' understanding.

In a context related to our study, Rouet, Vidal-Abarca, Bert-Erboul, and Millogo (2001) also found behavioral on-line evidence for integration processes being induced by comprehension tasks. They asked undergraduate students to search a 35-paragraph text in order to answer high- or low-level questions. High-level questions required the students to process and connect dispersed units of information in the text, whereas low-level questions only required the location of very specific textual units. The authors observed that whereas high-level questions promoted a review-and-integrate search pattern (i.e., the students reviewed more paragraphs of the text and answered the questions more times), low-level questions triggered a locate-and-memorize search pattern,

which was apparent in the search for specific pieces of information and in locating and memorizing them to give their answers. In another study, Cerdán, Vidal-Abarca, Martínez, Gilabert, and Gil (2007) found that those patterns were related to differences in learning from text, as we explain in the following section.

Mannes and Hoyes (1996) found behavioral correlates in terms of reading times for the integration process. They conducted three experiments in which students first studied an outline of an expository text which was either similar to or different from the text and then read the complete text, so that when students encountered an idea in the text that they had studied in the outline, the idea was reinstated and integrated with the text information. Mannes and Hoyes found that readers who were given the different outline built a richer domain representation of the text content compared with those who had read the similar outline, though the former task was more demanding than the latter. They also found on-line evidence for the advantage of the different outline. Students who had read this outline spent more time reading text ideas previously studied in the outline than those who had read the similar outline. This finding means that students in the different outline condition employed more cognitive resources to integrate the two sources than did those in the similar outline condition, and that this effort was responsible for the learning effects.

Goldman (2004) and other colleagues investigated students' strategies when reading multiple science documents that differed in reliability in order to explain a physical phenomenon (the eruption of Mount St. Helens). After reading, students wrote an essay explaining the phenomenon and ranked the reliability of the documents. It was found that good learners (i.e., those with significant gains from the text topic when comparing their knowledge before and after reading) produced more self-explanations on relevant documents and concentrated those self-explanations on regions of interest (i.e., text regions with information relevant to the criterial task) compared with poor learners. Moreover, poor learners seemed not to distinguish good from bad information in terms of the criterial task.

In summary, integration processes can be observed when recording reading behaviors with measurements such as rereading text information, slowing reading, or concentrating on information relevant to the task. These processes can be enhanced with tasks that make students connect different units of information. These tasks may require more effort, but in compensation they produce more learning than less demanding tasks. In the following section we explore this issue more deeply.

### Tasks to Promote Integration From Multiple Documents

Schmidt and Bjork (1992) reviewed some counterintuitive phenomena associated with the notions of training, performance, and learning. They presented evidence showing that experimental conditions that facilitate performance during training can be detrimental in the long term and, conversely, that manipulations that degrade the speed of acquisition or increase difficulty can nevertheless support the long-term goals of training. Therefore, according to these authors, learning should not be indexed by the improvements in skills across practice but should be measured as the level of posttraining performance. Additionally, making the process of learning more effortful may be helpful in terms of deep



learning. These two conclusions can also be applied to comprehension and learning from text.

Cerdán et al. (2007) had college students study a long science text by reading it and answering either high- or low-level questions with the text available; students' performance when answering the questions was recorded. Two days later the students were tested on two poststudy measurements that differed in the level of learning they were capturing (i.e., memory recall, inference, and problem solving). Students in the high-level questions condition scored higher than those in the low-level questions condition on the test measuring inferences and problem solving, though no differences were apparent on the memory recall test. Differences in the deep level of learning can be explained because of the search pattern when answering the two sorts of questions. Thus, whereas students answering high-level questions reviewed and connected more relevant units of information (i.e., a review-and-integrate search pattern), students answering low-level questions concentrated on locating isolated textual units (i.e., a locate-and-memorize search pattern). Hence, although high-level questions required the students to devote more effort to the task, they helped the students integrate text information, which is apparent when measuring deep but not superficial levels of comprehension.

In the context of the integration of information from different documents, Wiley and Voss (1999) manipulated readers' mental activity by instructing students to write arguments, narratives, summaries, or explanations when reading either a set of documents or the same information presented as a textbook chapter. In two experiments they found that writing arguments was more effective for making inferences and transforming information than the other conditions, but no differences were apparent when the test involved repeating information. Writing arguments made students establish more connections among ideas, particularly causal connections, which produced a more integrated situation model of the events described in the documents.

In summary, tasks that induce students to integrate ideas across a long text or a set of documents benefit making inferences and solving problems, even though such tasks require more effort. However, those benefits are not apparent with superficial learning, as this level of learning can be achieved with less cognitive effort. Thus, when measuring performance in the short term, no differences among more and less demanding tasks may appear, but these differences may be apparent with measures that capture a deep level of learning.

### Objectives and Hypothesis for This Study

Earlier we summarized our main purposes for this study. First, we wanted to deepen our understanding of the mental processes responsible for enabling integration from multiple sources and to find behavioral correlates for these processes. Second, we wanted to test the effectiveness of two sorts of tasks to induce integration and learning from multiple documents. The tasks were either (a) to write an essay on a broad or intertext question that involved selecting, processing, and combining a subset of information across the documents; or (b) to answer more narrow intratext questions covering the same information. These intratext questions induce less integration as they are answered individually and each is based on only one text.

To accomplish these goals, we selected three texts on the general topic of bacteria resistance to antibiotics from which students had to select, process, and combine a subset of information on a specific issue included in the documents, that is, mechanisms by which bacteria become resistant and how this resistance can be transmitted to other bacteria. Each of the texts partially presented this subtopic and also included irrelevant information. Thus, integration across relevant units of information coming from each document could take place. Only the student who extracted and combined the relevant units of information in each text could construct a coherent representation of the subtopic. In contrast, readers failing to extract and connect the relevant pieces of information in each text would not be able to construct an accurate and coherent representation of the main issue presented across the documents.

We hypothesized that the task of writing an essay on a broad or intertext question would be more effective to fostering the integration of relevant pieces of information across documents than answering the intratext questions. The former task would make students connect many of the ideas relevant to the task and to process them together, which would be apparent in both the on-line processing measures and the final learning result. However, no differences between the two sorts of tasks were predicted, whether on task performance or on memory recall, as the students would select the relevant information at the same level independently of the type of task.

We used the think-aloud method to obtain convergent on-line evidence for our experimental hypothesis. The think-aloud method has been demonstrated to reflect what is available in working memory, accessible to consciousness, and codable in language (Ericsson & Simon, 1980, 1993), and would therefore be indicative of which mental processes and content are responsible for how students perform and learn from a specific task. Indeed, there has been a growing use of this method in text comprehension research to obtain strategic data (i.e., Coté & Goldman, 1999; Goldman & Durán, 1988; Magliano, Trabasso, & Graesser, 1999; Pressley & Afflerbach, 1995; Trabasso & Magliano, 1996).

Thus, using a think-aloud procedure should provide valuable data on the processes and strategies that occur during comprehension, and the validity of these data would increase if convergence with other kinds of data were provided (Long & Bourq, 1996; Whitney & Budd, 1996). Indeed, converging evidence for the conclusions drawn from protocol data has been provided (Magliano & Millis, 2003; Suh & Trabasso, 1993; Trabasso & Magliano, 1996; Trabasso & Suh, 1993). Additionally, using the think-aloud methodology as one part of a converging evidence strategy was presented in Magliano and Graesser's (1991) three-pronged approach to comprehension research. Magliano et al. (1999) also used converging measures of think-aloud protocols and reading times. They correlated think-aloud content with silent-reading times and found a relationship between the increase or decrease in sentence-reading times and the type of text-processing strategy.

Nevertheless, an important controversy has arisen regarding the degree to which using a think-aloud procedure interferes with performance and learning when compared with silent-reading conditions. Ericsson and Simon (1980, 1993) argued that participants can generate verbalizations subordinated to task-driven cognitive processes without changing the sequence of their thoughts and slowing down only moderately due to the additional verbalization.



They insisted, providing substantial evidence based on research, that if participants are asked only to verbalize thoughts per se (*Type 1 and 2 verbalizations*), no interference effects will be apparent. Nevertheless, some experiments request participants to provide reasons, justifications, and elaborations (*Type 3 verbalizations*). In such cases, some studies found that verbalization affected performance (Ericsson & Simon, 1993), but the ways in which it did so were diverse, in some cases increasing performance and learning and in other cases hindering both.

This discrepancy is also present in text comprehension think-aloud studies with college students or adult readers. There is evidence that having students verbalize their thoughts while performing a task creates no difference compared to a silent-reading condition. Two studies with adult readers found no differences in comprehension between readers who stopped to give think-aloud comments and those who read silently (Crain-Thoreson, Lippman, & McClendon-Magnuson, 1997; Fletcher, 1986). Contrary to these results, there is also evidence that thinking aloud may hinder comprehension and learning from text among university students. For instance, Wade and Trathen (1989) studied the role of noting ideas in a text and its effects on recall. Of interest is the fact that they found that asking students to describe their strategies while studying reduced the amount of information students took notes on and had negative effects on recall. They concluded that verbal reporting during reading affects both the process and product of studying.

Magliano et al. (1999) conducted an experiment on how strategic processing during comprehension is affected by properties of the text and how different strategies (i.e., reading to explain, to associate, to predict, or simply to understand) can affect text retention. They compared silent and think-aloud conditions and found that a strategy to explain led to an increase in memory compared with a strategy to understand when reading was done silently but not when participants thought aloud. The authors argued that it is possible that thinking aloud eliminated the general benefits of reading to explain. Moreover, they pointed out that the process of describing the thoughts that occurred in a sentence may have strengthened memory representations for the explicit text. Indeed, story recall was better when participants thought aloud than when they read silently. In short, thinking aloud may improve memory for text because it requires a more conscious processing of text, though it eliminates the benefits of the student being induced to use an explanatory strategy (Magliano et al., 1999).

Following these results, we were interested in assessing the degree to which thinking aloud may become a source of cognitive overload (Sweller, 1994) in complex learning situations, which might distort processing or provide convergent evidence of the integration processes in conjunction with behavioral on-line measures. It might be that thinking aloud, as hypothesized by Ericsson and Simon (1980, 1993), does not interfere with learning from a task and perhaps creates only a slight increase in experimental time because of the need to think aloud and perform the task simultaneously. However, thinking aloud in conjunction with reading multiple texts and solving different tasks may overload a student's cognitive resources, resulting in decreased performance and/or learning. In terms of processing patterns, thinking aloud may affect the reading of texts, inducing greater concentration on explicit textual information, which might decrease the use of deep processing strategies such as in Magliano et al.'s (1999) study.

## Method

### Participants

Fifty-six undergraduate students with a mean age of 20 years who were enrolled in a psychology program at the University of Valencia, Valencia, Spain, participated in the experiment for course credit. Some participants were excluded because of problems with the on-line registration process. They were distributed into four conditions following a matching procedure based on their previous background knowledge level (see *Procedure* section). Half of the participants were assigned to the intertext task, whereas the other half answered intratext questions. Within each group, approximately half of the participants thought aloud while doing the task (i.e.,  $n_s = 11$  and 12 for the intertext and intratext questions, respectively), whereas the rest did the task under silent conditions (i.e.,  $n_s = 14$  and 13 for the intertext and intratext questions, respectively).

### Materials

*Documents and tasks.* Three texts from trustworthy Web sites (e.g., the National Institutes of Health) were selected according to length, comprehensibility, and content-relevance criteria. Thus, we guaranteed that texts both were written by a specialist and came from reliable sources. The texts presented information to be understood by a nonacademic in biology. Nonetheless, the content was not suitable for the general public, but rather for university-level students used to handling scientific texts and with general background knowledge in biology acquired previously.

Our main interest was to create a situation of learning from multiple texts based on complementary relationships across sources. This situation implies that, in order for a reader to wholly understand the specific issue (i.e., mechanisms by which bacteria become resistant and how resistance can be transmitted to other bacteria), he or she would need to process carefully the relevant pieces of information coming from each of the three texts and combine them. In other words, none of the sources gave the whole perspective on the issue; instead, the reader needed to select relevant units of information across sources and integrate them into a higher order representation going beyond the individual texts. The three texts (i.e., *Anti-Bacteria Resistance*, *New Perspectives on Bacteria Resistance*, and *Genetics of Bacteria Resistance*) were 684, 658, and 390 words long, respectively.

After conducting a detailed content analysis of the three texts to clearly specify which units of information were relevant to the specific issue, we selected 22 units as relevant (see Table 1). We (Raquel Cerdán and Eduardo Vidal-Abarca) performed content analysis and selected units, reaching an interrater agreement of more than 90%. The 22 idea units were grouped into four main content areas. These areas corresponded to the main subtopics included in the specific issue put to students, that is, description of bacteria, biological resistance mechanism, the transfer of bacteria resistance from some bacteria to others, and resistance responses. The units were differentially present in the three documents (see Table 1). Therefore, any student wishing to construct an integrated mental model would have needed to inspect all sources, select the relevant units of information, and integrate them. Though many units were located in different documents, five units repeated in two different texts (i.e., Ideas 7–9, 15, and 16; see Table 1). This

Table 1  
*Distribution of Contents Across Texts*

Idea	Content	Text	Paragraph
Area 1. Description of bacteria			
1	Bacteria are single-celled organisms with a reduced number of genes.	R	2-1
2	They multiply quickly.	R	2-1
3	They are highly capable of adapting to any kind of environment.	R	2-1
4	This is a key factor for the development of resistance to antibiotics.	R	2-1
Area 2. Biological resistance mechanisms			
5	When antibiotics appeared, many bacteria developed resistance mechanisms.	P	1-3
6	There are two resistance mechanisms.	P	1-3
7	Genetic mutation.	P, G	P 1-3, G 2-1
8	The transfer of resistance genes from one bacterium to other bacteria (plasmids)	P, G	P 1-3, G 2-1
9	Genetic mutation implies changes in the genetic information of bacteria.	P, G	P 1-3, G 2-1
10	These changes prevent them from responding to antibiotics.	P	1-3
11	Plasmids are pieces of extra-chromosomal DNA.	P	1-3
12	Plasmids produce mechanisms resistant to antibiotics.	P	1-3
13	Plasmids may produce resistance to more than one antibiotic. This is called multiresistance.	G	2-1
Area 3. The impact and transfer of resistance			
14	Even one random mutation can have a large impact due to bacteria's quick multiplication rate.	R	2-1
15	Resistance can be transmitted to other generations of the same bacteria.	P, R	P 1-3, R 2-1
16	Resistance can also be transmitted to other generations of bacteria not related to the original ones.	P, R	P 1-3, R 2-1
Area 4. Resistance responses			
17	Bacteria respond in three different ways to the lethal effects of antibiotics.	G	1-3
18	They can modify the antibiotic's chemistry.	G	1-3
19	They can degrade the antibiotic.	G	1-3
20	They can modify the target of the antibiotic in the bacteria.	G	1-3
21	They can stop antibiotics from penetrating.	G	1-3
22	They can expel the antibiotic.	G	1-3

*Note.* Area refers to the main content across texts. Paragraph refers to Read&Answer text divisions. R = *Anti-Bacteria Resistance*; P = *New Perspectives on Bacteria Resistance*; G = *Genetics of Bacteria Resistance*.

repetition of information units was not verbatim but at the level of meaning.

The four main content areas were the basis of the design of the two sorts of tasks. Writing the essay on the intertext question required the students to take into account the 22 units of information (see Table 1) and discard irrelevant ideas. Complementarily, four intratext questions that directly matched the four areas in the table of contents were formulated (see the Appendix for the two types of questions). Therefore, answering the intertext question promoted greater integration of the units as it required students to connect information from the three documents, although this task might have been more demanding than the other question. However, less integration was needed to answer the four intratext questions, as their answers were found only in one document, thus not requiring students to make intertext connections.

*Read&Answer software* (Vidal-Abarca & Martínez, 2002). Texts and tasks were presented on a computer screen using the software Read&Answer. The software presents readers with a full screen of text. All text except the segment currently selected by the reader is masked. Readers unmask a segment by clicking on it; when they unmask another segment, the first segment is remasked. Thus, only one segment is visible at a time. Readers can reread the segments in any order they choose (see Figure 1A). A segment can

be a word, sentence, or paragraph, depending on the unit the researcher establishes. In this study, a text segment was operationalized as a paragraph, which, as described below, was the level of text at which measures were taken in the Read&Answer system. Each text was presented in a maximum of two pages. The first and the second texts contained seven paragraphs each, and the last text contained only five. Each paragraph had a mean number of words of 90. Relevant units of information were located in four Read&Answer paragraphs, whereas the rest of the paragraphs were irrelevant to the questions in the tasks (see Table 1).

Read&Answer also presents the reader with a question screen divided into two parts, the upper part for the question and the lower part for the answer. This question screen can be accessed through a button on the navigation toolbar from any text that the student may be reading at a given moment, thus allowing flexibility in the question-answering process. The user clicks on each part of the question screen to either read the question or write in the answer box. A simple interface allows the reader to move from one question to another and from the question screen to the text screen, and vice versa (see Figure 1B).

Read&Answer automatically generates three outputs. The first is a list of all segments active at any given moment, which is sequentially ordered to follow the student's performance in the



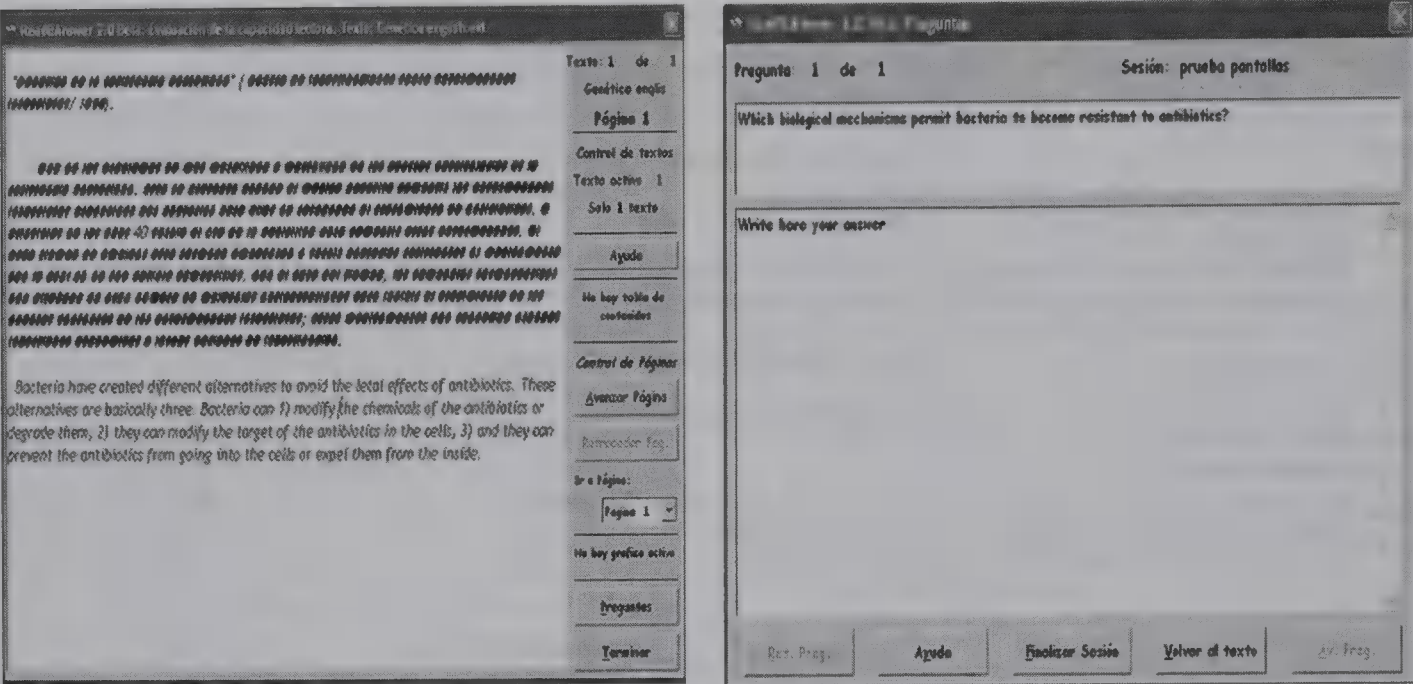


Figure 1. Read&Answer text (A) and question (B) screens.

experimental session. The software also records the length of time each segment is active. A text paragraph, a specific question, and the answer to every question are all segments. Thus, every action the reader undertakes, whether reading a text paragraph, reading a question, rereading a paragraph, or writing an answer, is recorded and included in the sequential list. The second output is a summary of the reader's behavior when he or she reads the text and answers the questions, including computed data such as the reading rate per word for the textual paragraphs, which the system calculates by dividing the total time to read that paragraph by the total number of words in that paragraph. The third output is the record of the reader's answers to each question. These three outputs allowed us to track students' strategic behavior when reading multiple sources on-line.

*Processing measures.* To measure students' processing of information, we took a number of on-line measures using the Read&Answer tracking system, assuming that they would contribute to explaining the product results. We first measured time on task. Time on task in seconds indicated how much time students needed to perform the reading and question-answering task. We also captured the students' processing of both relevant and non-relevant units of information by measuring the mean time per word that students in the different conditions spent reading the two sorts of units of information.

Finally, we measured integration of information by counting nonconsecutive readings of relevant units of information, which indicated an effort to connect and integrate the two paragraphs. We also measured the single-unit pattern of processing by counting the number of times a student, after reading a paragraph with relevant information, performed another action (e.g., reading a question, writing an answer, or going to a paragraph with nonrelevant information) and then returned to the paragraph with the previously read relevant information. This behavior indicated the processing of text paragraphs in isolation from other paragraphs in the three texts.

Students' think-aloud protocols were also analyzed to obtain complementary evidence for processing. Students' verbalizations were transcribed and divided into idea units, that is, sentences with a subject and a verb. Each idea unit was coded according to the following categories: (a) *task*: verbalizations associated with the process of performing the task or the use of the software (e.g., "These are the instructions, aren't they? I read them aloud. Okay, I will start by reading the questions."); (b) *search*: verbalizations regarding searching and locating information in the texts, regardless of whether it was relevant or irrelevant (e.g., "Okay, I think this was located in Text 3. I will reread Text 3 to find the answer."); (c) *relevant-unit understanding*: verbalizations about relevant units of information that indicated deep comprehension, including inferences, summaries, elaborations, and explanations (e.g., a student said the following when she was trying to understand the mechanisms by which bacteria produce resistance to antibiotics: "Here they more or less give a different point of view of how resistance is produced; given that they are mono-cells and with a tiny amount of genes, by mutation of these genes, they can create resistance, which is transmitted to their descendants"); (d) *non-relevant-unit understanding*: deep comprehension verbalizations, but referring to irrelevant units of information; (e) *superficial understanding*: verbalizations indicating paraphrases and irrelevant associations, no matter if they referred to relevant or irrelevant units of information; (f) *writing self-regulation*: verbalizations reflecting self-regulation of the answering process (e.g., "I go to Question 3, because I don't know what to answer in this one"; "I think this one is correct this way"); (g) *superficial writing*: verbalizations indicating mainly how students manage orthography in the writing process (e.g., "I am not sure if I should add a period at the end of this sentence"). Verbal protocols were coded by two experimenters according to these categories. Coding was done following the procedure explained previously: cooperative scoring for training, coding approximately one third of protocols

separately, but checking interrater agreement until reaching 95% agreement; then Raquel Cerdán coded the rest of the protocols.

*Control measures.* We used three control measures to make sure the experimental groups did not differ in measures that could contaminate subsequent results. The first control measure was prior background knowledge. As students were expected to learn selected content on bacteria resistance to antibiotics from three different texts, and differences in previous background knowledge may have interfered with our results, we created a measure to control for this possible effect. In general, we expected students to have similar levels of basic knowledge of general biological mechanisms from earlier instruction in school. Additionally, we expected them not to be specialists in bacteria resistance mechanisms. This way, they would be a suitable sample for learning this specific knowledge in our experiment.

To ascertain students' level of knowledge of biology, we administered a background knowledge test that consisted of 15 true/false items on the general domain of biology relevant for the texts, although 3 of the 15 items directly assessed specific content on bacteria resistance that students would encounter in the texts. Of these items, 8 were true and 7 false, and the test was scored up to 15 points. An example of a false general domain item was the following: "Bacteria are multi-celled organisms." This contrasted with an example of a true topic-specific item: "Bacteria resistance is favored by the misuse of antibiotics both in humans and animals." Reliability analyses were conducted with an independent sample of 45 psychology students. Cronbach's alpha was significant, with a value of .97.

The second control measure was lexical access. A Spanish test consisting of a list of 40 words and a complementary list of 40 pseudowords was presented to the students, and we registered the time students took and the mistakes they made in reading aloud both lists (Ramos & Cuetos, 2000). Third, we controlled for keyboarding skills, as students used a computer to answer the questions. Participants were given 2 min to type a 146-word text into the computer. Then we counted the number of words they were able to type.

*Tests and scoring.* To assess the effectiveness of the two tasks at different levels of learning, we designed three measures: (a) performance on the tasks, (b) a sentence verification task, and (c) a transfer task. The first captured how well students did on either the intertext task or the four intratext questions, whereas the other measures were applied after the students finished the tasks. One measured superficial understanding, whereas the other measured deep learning.

To measure performance on the tasks, we counted whether the 22 target ideas that students were expected to include in their answers were present, as these target ideas were shared in both conditions. Thus, a student performing the intertext task perfectly would obtain 22 points, 1 for each idea included in the answer. Performance scores were transformed into percentages to facilitate interpretation. The scoring procedure was conducted by two experimenters. To guarantee equality in criteria, at least two question sets were scored cooperatively and compared between scorers. Discrepancies would be discussed and resolved. Then, new question sets were scored separately and rater agreement was checked after scoring each protocol until raters agreed on at least 95% of decisions (this was reached after scoring approximately one third

of the protocols). The rest of the protocols were scored by Raquel Cerdán alone.

The sentence verification test consisted of 18 items that directly matched the relevant units of information that students should have gone through in the experimental session. Each of the items included either one or two of the units present in the basic table of contents across texts. Examples of items are the following: "Plasmids are DNA pieces" (false) and "Plasmids produce resistance mechanisms against antibiotics" (true). These items correspond to Ideas 11 and 12 in the table of contents, respectively. Reliability analyses were again conducted with an independent sample of 45 psychology students. Cronbach's alpha was significant, with a value of .94.

The transfer test consisted of a practical case in which students had to apply to a new situation their new knowledge on bacteria resistance to antibiotics. Because of the need to transfer coherent knowledge acquired from reading and to connect the relevant paragraphs of information, we assumed the transfer test would measure integrated knowledge. Thus, students were presented with the following practical situation: "Imagine you have a sore throat and you go to the doctor for treatment. The doctor informs you that the cause of your illness is a bacterial infection. You should take antibiotics for an entire week. You start your treatment, following the doctor's advice, but after 3 days you feel perfectly well and decide to stop taking antibiotics." Based on this situation, students were required to answer the following questions: (a) "Do you think the antibiotic has been successful in destroying all the bacteria responsible for your illness, and why?"; (b) "What explanation can you give?"; (c) "What further implications would occur because you stopped taking the antibiotics?" The test was scored up to 9 points, and a percentage of success was obtained for each student.

The criteria to score the three questions were the following. For the first question students were expected to answer no (1 point). They should reason that when antibiotics are used improperly, bacteria resistance may appear (1 point), and hence these antibiotics would be unable to destroy all of the bacteria responsible for the illness (1 point). For the second question students should have given a biological explanation for the reasons given for the first question: Bacteria may have become resistant to antibiotics (1 point), either because genetic mutation had taken place (1 point) or because DNA pieces producing bacteria resistance had been created (1 point). Finally, a complete answer to the third question included the following causal chain: Bacterial resistance can easily be transmitted to other bacteria (1 point). This transmission is favored by the bacteria's high multiplication rate (1 point). Thus, an individual action (e.g., to stop taking antibiotics) may affect the long-term effectiveness of a specific antibiotic (1 point). The scoring system was the same as that explained previously, involving a collaborative training phase, then an independent scoring phase in which interrater agreement was checked until reaching 95% agreement (which was also achieved after scoring approximately 25% of the protocols). At that point the rest of the protocols were scored by Raquel Cerdán.

### *Procedure*

Students were assessed individually during two sessions on 2 different days. In the first session, students were tested on the three control measures (i.e., previous background knowledge, lexical



access, and keyboarding skill). Students were matched to one of the four experimental conditions (Think Aloud  $\times$  Task), taking into account their scores on the previous background knowledge test so that the mean score of students in the four conditions would be equivalent. Correction of the previous background knowledge test and assignment to each of the conditions was done while a specially trained assistant tested the students on the lexical access and keyboarding tasks in order to match the students to the four experimental conditions. Afterward, students were trained to use the Read&Answer software with a task similar to the one they had to do in the second session. Specifically, students were first told to read the task, either the intertext essay or the four intratext questions, and then move about freely in the software, reading the texts as they wished and going back to their task whenever they wanted.

In Session Two, students did the experimental task. They first read either the intertext task or the four intratext questions, and the titles of the documents. Thus, they knew the sort of question they had to answer in advance. Then students were free to read and answer their respective questions at will, being free to read and reread the documents when doing their task. After finishing, they were tested first on the sentence verification test and then on the transfer test. These two final tests were performed with paper and pencil and without a time limit, though students used a similar amount of time for the two tasks (approximately half an hour).

The procedure was slightly different for the students in the two think-aloud groups. In the first session the training also emphasized that students should read aloud the textual paragraphs and simultaneously say whatever came to their minds during this performance. Students were invited to sit in front of the computer and were given a headphone with a microphone to record their verbalizations. We told them to say aloud whatever crossed their minds while reading the texts and answering the questions, but without making an extra effort to verbalize, as they should say aloud only naturally appearing thoughts that occurred while performing the task. We indicated to students that they were expected to verbalize (a) after reading each of the text paragraphs of the documents; (b) while moving through the software (i.e., changing from one document to another and going from the text screen to the question-answering screen); and (c) before, during, or after reading the questions and writing the answers. Then one of the authors modeled the think-aloud procedure for 5 min using different types of verbalizations, such as those implying repetition, paraphrases, or inferences.

Results

A number of  $2 \times 2$  analyses of variance (ANOVAs), with the two main variables task (intertext task vs. intratext questions) and think aloud (yes vs. no), were conducted for most of the dependent measures. We first briefly present results for the control measures. Then we explain the product measures and those capturing the processing. After that, correlations between product and processing measures are presented.

Control Measurements

We found no significant differences for any of the control measurements: previous background knowledge, lexical access skill, or, finally, keyboarding skill (see Table 2). Therefore, we had strong assurance that the experimental conditions did not differ in measures that could have interfered in subsequent analyses. Of interest is the fact that the mean score in all groups for background knowledge was 9.5, out of a maximum of 15 points. This finding indicates that participants had a relatively high level of general domain knowledge that would have allowed the learning of more specific issues, such as bacteria resistance.

Product Measurements

*Performance.* Neither the type of task nor the think-aloud conditions generated significant main effects on performance,  $F(1, 46) = 0.10, p = .79$ ; and  $F(1, 46) = 1.13, p = .48$ , for task and think-aloud, respectively, though the interaction between them was marginally significant,  $F(1, 46) = 3.12, p = .08, \eta^2 = .06$ . Thus, as can be seen in Table 3, on intratext questions participants scored higher when they thought aloud than when they did not, whereas thinking aloud did not make any difference for the intertext task. This interaction effect could be interpreted as a *textual-focusing effect* of thinking aloud, which is added to the local processing effects of intratext questions. Thinking aloud could have strengthened the student's maintenance of textual information in short-term memory to a greater extent than the non-think-aloud condition in a situation in which local processing was enhanced by answering intratext questions. It could have made students who thought aloud and answered intratext questions increase their performance scores.

*Sentence verification test.* Neither the type of task effect nor the think-aloud conditions produced significant results on sentence

Table 2  
Control Measures

Group and task	Background knowledge		Time words		Time pseudowords		Mistakes words		Mistakes pseudowords		Keyboarding	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Think aloud												
Intertext	9.55	2.98	31.27	6.62	44.45	7.63	0.00	0.00	1.36	1.86	67.27	16.54
Intratext	10.25	2.01	28.58	4.21	43.92	7.33	0.08	0.29	1.00	1.28	64.00	13.02
Non-think-aloud												
Intertext	9.29	2.92	29.00	7.83	41.86	8.10	0.00	0.00	1.50	1.61	61.85	21.05
Intratext	9.08	2.22	30.00	4.98	41.62	6.71	0.15	0.38	2.00	1.83	62.53	14.44

Table 3  
Means and Standard Deviations of the Three Product Measurements

Group and task	Performance		Sentence verification		Transfer	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Think aloud						
Intertext	47.52	25.98	51.01	30.10	40.90	16.34
Intratext	63.25	25.45	51.85	26.52	33.79	12.63
Non-think-aloud						
Intertext	46.75	22.58	53.57	22.70	53.96	15.47
Intratext	38.81	20.54	41.02	23.62	40.59	11.19

Note. Data are in percentages.

verification,  $F(1, 46) = 0.64$ ,  $p = .42$ ,  $\eta^2 = .14$ ;  $F(1, 46) = 0.32$ ,  $p = .57$ ,  $\eta^2 = .00$ ; and  $F(1, 46) = 0.84$ ,  $p = .36$ ,  $\eta^2 = .01$ , for task, think-aloud, and the interaction effects, respectively (see Table 3). Thus, as predicted, the two sorts of tasks were equally effective for identifying and understanding the basic ideas that students should have considered to do their task.

*Transfer test.* The two main variables produced significant effects (see Table 3). Thus, participants performing the intertext task scored higher ( $M = 48.22$ ,  $SD = 16.87$ ) than participants answering intratext questions ( $M = 37.33$ ,  $SD = 12.16$ ),  $F(1, 46) = 6.21$ ,  $p < .05$ ,  $\eta^2 = .12$ , which was predicted in our first hypothesis. Additionally, participants who did not think aloud scored higher ( $M = 47.53$ ,  $SD = 14.96$ ) than those who thought aloud ( $M = 37.19$ ,  $SD = 14.96$ ),  $F(1, 46) = 6.61$ ,  $p < .05$ ,  $\eta^2 = .11$ .

Therefore, two main results are apparent. First, answering the intertext questions, that is, writing a global essay, produces deeper learning than answering intratext questions, though no differences appear at the level of superficial comprehension, that is, performance on the tasks and the sentence verification task. These results were in agreement with our predictions and should have been related to the processing measures that we are explaining. Second, thinking aloud seems to hinder deep learning from multiple documents when the task is to select and integrate a subset of information across the documents, but it also seems to strengthen the local processing of answering intratext questions. Processing measures should contribute to explaining these results.

### Processing Measures

*Time on task.* To obtain the first on-line evidence of the task effects, we globally counted the number of seconds students spent performing the reading and answering tasks. This measure was an indicator of how resource demanding the task was and how thinking aloud may have affected the experimental process. Thus, we conducted a  $2 \times 2$  ANOVA with independent variables task (intertext task vs. intratext questions) and think aloud (yes vs. no).

The type of task did not produce significant effects ( $M = 2,291.48$  s,  $SD = 886.05$ ; and  $M = 2,615.28$  s,  $SD = 898.19$ , for the intertext task and intratext questions, respectively). However, thinking aloud did produce differences, as students who thought aloud spent more time on the whole experiment ( $M = 2,906.64$  s,  $SD = 856.39$ ) than those who did not ( $M = 2,067.27$  s,  $SD = 750.11$ ),  $F(1, 46) = 14.19$ ,  $p < .05$ ,  $\eta^2 = .23$ . The interaction between the two variables was also significant. Participants answering intratext questions needed more time for the experiment ( $M = 3,271.81$  s,  $SD = 655.09$ ) than those answering the intertext task, but only for the think-aloud condition,  $F(1, 46) = 3.99$ ,  $p < .05$ ,  $\eta^2 = .08$  (see Table 4). Thus, on the whole, thinking aloud increased the time needed to perform the tasks in comparison with participants performing silently, and the effect was stronger for the intratext questions task. Therefore, processing data support the textual focusing effect we mentioned when performance results were analyzed. Apparently, thinking aloud strengthened the students' maintenance of text information in short-term memory, enhancing the local processing caused by answering intratext questions, which provoked the high performance of students in this condition.

*Processing relevant vs. nonrelevant units of information.* The benefits of the intertext task found in the transfer test should have been apparent on the processing of the relevant information when doing the tasks. To examine this issue we analyzed the time per word of the students' processing when reading the text paragraphs with the 22 relevant units of information and when reading the rest of information, that is, the nonrelevant information. A  $2 \times 2 \times 2$  ANOVA was conducted with task and think aloud as between-subjects variables and type of information (relevant vs. irrelevant) as a within-subjects variable. The task variable produced significant differences, as intertext task participants read textual information more slowly, that is, at a higher time per word rate ( $M = 0.19$ ,  $SD = 0.06$ ), than students answering intratext questions ( $M = 0.15$ ,  $SD = 0.05$ ),  $F(1, 46) = 4.99$ ,  $p < .05$ ,  $\eta^2 = .09$ . This

Table 4  
Means and Standard Deviations of the Five Processing Measurements

Group and task	Time on task		Time per word relevant		Time per word nonrelevant		Integration		Single-unit processing	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Think aloud										
Intertext	2,508.27	898.24	0:20	0.07	0.22	0.05	4.55	2.018	12.00	10.45
Intratext	3,271.81	655.09	0.18	0.05	0.20	0.09	1.83	0.937	26.50	12.602
Non-think-aloud										
Intertext	2,121.14	870.47	0.21	0.09	0.13	0.05	6.71	5.622	10.07	6.765
Intratext	2,009.25	625.48	0.11	0.03	0.12	0.04	1.00	0.91	26.69	6.921



finding means that the intertext task induced a more detailed processing of documents than the intratext question task. The think-aloud variable also produced significant differences. Hence, think-aloud participants read globally more slowly ( $M = 0.20$ ,  $SD = 0.06$ ) than non-think-aloud students ( $M = 0.14$ ,  $SD = 0.05$ ),  $F(1, 46) = 12.09$ ,  $p < .05$ ,  $\eta^2 = .20$ , which seems logical as the time devoted to thinking aloud was counted as time for processing.

Three interesting interaction effects were apparent. First, the interaction between task and type of information was significant,  $F(1, 46) = 4.98$ ,  $p < .05$ ,  $\eta^2 = .09$  (see Figure 2). Thus, whereas intertext task participants read relevant information more slowly than irrelevant information, intratext question students read both sorts of information at a similar rate of speed. This may explain why answering an intertext task was more beneficial for deep learning than was answering intratext questions. Second, the interaction between thinking aloud and type of information was also significant,  $F(1, 46) = 7.43$ ,  $p < .05$ ,  $\eta^2 = .13$ . Relevant and nonrelevant information was read at a similar speed in the think-aloud condition ( $M = 0.19$ ,  $SD = 0.06$ ;  $M = 0.21$ ,  $SD = 0.07$ , for relevant and irrelevant information, respectively), whereas in the non-think-aloud condition, relevant information was read more slowly than nonrelevant information ( $M = 0.16$ ,  $SD = 0.06$ ;  $M = 0.12$ ,  $SD = 0.04$ , for relevant and irrelevant information, respectively). It seems that thinking aloud induced students to process all types of information at the same level, regardless of its relevance, which may explain why thinking aloud also hindered the transfer test scores.

Finally, a triple interaction among task, thinking aloud, and type of information was found,  $F(1, 46) = 5.79$ ,  $p < .05$ ,  $\eta^2 = .11$  (see Table 4). If we consider the relevant information, when students did not think aloud, those who performed the intertext task read more slowly than students who answered intratext questions, whereas no differences between the two sorts of tasks were found when students did the think-aloud task. However, when irrelevant information is considered, no differences were apparent between the two sorts of tasks, either when students thought aloud or when they did not, and students who did not think aloud read this irrelevant information more quickly than those who did think aloud. This finding indicates that the benefit of performing the intertext task over the intratext questions concentrates on the students' processing of the relevant information when they did not

think aloud. Therefore, the negative impact of thinking aloud on processing was confirmed.

*Integration versus single-unit processing of relevant units of information.* Integration of information should be apparent when a student reading a text paragraph with relevant information jumps to a nonconsecutive paragraph that also contains relevant units of information, no matter if in between the reader rereads a question paragraph or makes a quick visit to an irrelevant paragraph in the search process. One can assume that the reader is making an effort to connect and integrate the two paragraphs with relevant information. For this connection between two relevant paragraphs to occur, the in-between reading times (i.e., reading other irrelevant paragraphs or reading the question) should be less than the reading time of the two paragraphs being connected. Hence, one can assume a connection between two relevant paragraphs is taking place when two criteria are met: (a) the reading of the two relevant paragraphs is consecutive, and (b) the reading times of the in-between information are less than the reading times of the two relevant paragraphs being connected.

A completely different behavior occurs when a student reading a paragraph with relevant information performs another action, such as reading a question, writing an answer, or going to a paragraph with irrelevant information, and then goes back to the paragraph with relevant information previously read for further slow processing. This behavior reflects the processing of a specific unit of information without integration with other units. It should be noted that many students returned to the same unit repeatedly, but we counted these visits as one instance of processing, as they referred to the same unit.

We call the first instance *integration processing*, as it indicated an effort to connect different relevant paragraphs of the text, and the second *single-unit processing*, as it reflected the processing of the textual paragraphs in isolation to the other paragraphs. We counted the number of times a student either integrated information or showed the single-unit processing strategy just mentioned by applying the criteria just explained.

Two  $2 \times 2$  ANOVAs were conducted with task and think aloud as independent variables, and integration and single-unit processing as dependent variables. Only the task variable yielded significant results for both dependent measures (see Table 4). Students in the intertext question task, as expected, integrated information to a greater extent ( $M = 5.76$ ,  $SD = 4.47$ ) than did intratext question students ( $M = 1.40$ ,  $SD = 1.00$ ),  $F(1, 46) = 21.48$ ,  $p < .05$ ,  $\eta^2 = .31$ . The opposite trend was found for the single-unit processing measure, as intratext question students showed more single-unit processing ( $M = 26.60$ ,  $SD = 9.83$ ) than did intertext question participants ( $M = 10.92$ ,  $SD = 8.44$ ),  $F(1, 46) = 34.45$ ,  $p < .05$ ,  $\eta^2 = .42$ . The first result contributes to explaining why performing the intertext task produced higher scores on the transfer test than did answering intratext questions, as jumping from one relevant information paragraph to another contributes to the integration of the two paragraphs and consequently to deep understanding. However, the repeated processing of one paragraph does not contribute to integration but indicates the processing of one isolated unit of information.

*Analysis of think-aloud protocols.* For each student we obtained a cumulative score in each of the seven think-aloud categories mentioned in the *Method* section (i.e., task, search,

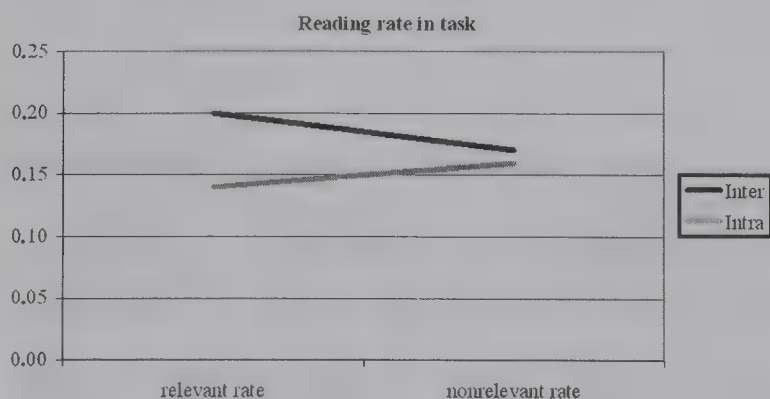


Figure 2. Interaction effect between task and type of information. Mean relevant and nonrelevant time per word rates for intertext and intratext questions.

Table 5  
*Think-Aloud Categories*

Task	Task		Search		Relevant-unit understanding		Nonrelevant-unit understanding		Superficial text processing		Writing self-regulation		Superficial writing	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Intertext	93.36	37.77	25.81	29.12	15.18	12.57	9.72	10.55	15.72	6.00	55.18	25.63	2.72	2.96
Intrertext	109.91	33.45	28.41	13.02	6.25	6.67	14.50	26.35	24.91	37.21	52.41	23.93	8.75	6.21

Note. Data are frequencies.

relevant-unit and nonrelevant-unit understanding, superficial understanding, writing self-regulation, and superficial writing) based on the number of times each category appeared in a verbal protocol (see Table 5). With these categories as dependent variables, we conducted one-way ANOVAs, with task (intertext task vs. intratext questions) as the independent variable. Only two analyses yielded significant results. First, students performing the intertext task ( $M = 15.18$ ,  $SD = 12.57$ ) verbalized significantly more on understanding relevant information than those who answered intratext questions ( $M = 6.25$ ,  $SD = 6.67$ ),  $F(1, 21) = 4.64$ ,  $p < .05$ ,  $\eta^2 = .21$ . Second, students answering the intertext task produced fewer superficial writing comments ( $M = 2.72$ ,  $SD = 2.96$ ) than students answering intratext questions ( $M = 8.75$ ,  $SD = 6.21$ ),  $F(1, 21) = 8.53$ ,  $p < .05$ ,  $\eta^2 = .46$ . These two results provide convergent evidence for our claim that the intertext task promoted a deeper processing of the multiple sources in comparison to the intratext questions.

### *Correlations Between Product and Processing Measurements*

To gather additional evidence to support the relationship between students' processing and the results obtained in the three product measures, we obtained the correlation between the processing and product measures (see Table 6). Performance on task was positively related to single-unit processing ( $r = .356$ ,  $p < .05$ ) and negatively related to time per word reading relevant units of information ( $r = -.332$ ,  $p < .05$ ). The opposite relationship was seen for scores on the sentence verification test, which were negatively related to single-unit pro-

cessing scores ( $r = -.352$ ,  $p < .05$ ) and positively related to time per word reading relevant units of information ( $r = .348$ ,  $p < .05$ ). Transfer test scores were also negatively related to single-unit processing ( $r = -.300$ ,  $p < .05$ ) but strongly related to integration behavior ( $r = .454$ ,  $p < .01$ ).

We submitted the processing variables to three stepwise multiple regression analyses to determine the best model for predicting the three product measurements. As can be seen in Table 7, only single-unit processing accounted for almost 11% of the variance in the performance scores, and no other processing measure accounted for a significant amount of variance above that mentioned. Single-unit processing was also the best and single predictor of variance (10.6%) in the sentence verification scores, though the beta coefficient was negative. In terms of the transfer test, the scores on integration behavior accounted for 19% of the variance, and no other processing measure made a significant contribution. Thus, the two variables related to the process of integration (i.e., single-unit processing and integration behaviors) were the best predictors of the three product measures, though their role was very different depending on the specific product measure we considered. Single-unit processing seemed to indicate very superficial processing characterized by searching for, locating, and memorizing information. This behavior was sufficient to succeed in performing the tasks during training but was negatively related to truly understanding the ideas found in the documents. Integration predicted transfer scores, the deepest comprehension measurement. It should also be noted that performance on the task was not related to the other two measures (i.e., sentence verification and transfer tests), whereas these two measures were closely interrelated ( $r = .357$ ,  $p < .05$ ).

Table 6  
*Correlations Between Pairs of Measurements*

Measure	Performance	Sentence verification	Transfer	Time on task	Time per word relevant	Time per word nonrelevant	Integration	Single-unit processing
Performance	—	.154	-.003	.261	-.332*	-.127	.002	.356*
Sentence verification		—	.357*	-.119	.348*	.126	.141	-.352*
Transfer			—	-.168	.144	-.211	.454**	-.300*
Time on task				—	.019	.234	.060	.441**
Time per word relevant					—	.587**	.195	-.607**
Time per word nonrelevant						—	-.225	-.260
Integration							—	-.421**
Single-unit processing								—

\*  $p \leq .05$ . \*\*  $p \leq .01$ .



Table 7  
*Multiple Regression Predicting the Three Product Measures*

Predicted variable	Predictor	$R^2$	$\beta$	$t$	$p$
Performance	Single-unit processing	.108	.356	2.636	.011
Sentence verification	Single-unit processing	.106	-.352	-2.609	.012
Transfer test	Integration processing	.190	.454	3.533	.001

## Discussion

Our main objectives for the present experiment were (a) to deepen our understanding of the mental processes involved in integrating information from multiple documents and to find behavioral correlates for these processes; and (b) to test the effectiveness of two sorts of tasks to induce integration and deep learning from multiple documents, that is, either writing an intertext essay or answering intratext questions, both drawing on the same text information. Accordingly, we hypothesized that students writing the essay or intertext question would outperform students answering the intratext questions on deep learning measures, but not on superficial comprehension, and that this differential effect could be explained from on-line processing measures while doing the tasks. Additionally, we wanted to explore possible interference effects of thinking aloud while doing the tasks, given the debate about this.

Writing an intertext essay, which according to our design should have made students integrate the target relevant paragraphs across the three texts, was indeed the most effective task for fostering deep learning, as evidenced in the transfer test. Thus, students who wrote the essay were better able to apply their new knowledge on bacteria resistance to antibiotics to a new situation than were those who answered intratext questions. However, when considering performance scores on the tasks or results at the level of identifying and understanding isolated relevant ideas, there were no differences between writing an intertext essay and answering intratext questions. Thus, measures of performance and superficial comprehension are not coincident with measures of deeper comprehension and learning (Schmidt & Bjork, 1992; Wiley & Voss, 1999).

Our analysis of students' processes based on behavioral data allows us to explain these results. The instruction to write an essay on an intertext question made students read the relevant information more slowly than irrelevant information, whereas no differences between the processing of the two types of information were found when students answered the intratext questions. In addition, students who wrote the essay showed more integration behavior than those who answered intratext questions (such behavior defined as reading a text paragraph with relevant information, then jumping to another nonconsecutive paragraph that also contains relevant units of information, whether or not the reader performs other actions in between these tasks). However, single-unit processing behaviors were more frequent in answering intratext questions than in writing the essay on the intertext question. We defined these behaviors as reading a paragraph with relevant information, then doing other actions, and finally going back to the paragraph with the relevant information previously read for slow processing.

Therefore, writing an essay on a broad and global issue makes students concentrate on the information relevant to the essay topic

and integrate all of the information even though it may be a highly demanding task. Indeed, connecting different paragraphs of information coming from different sources seems to be more demanding for the student than inspecting a single text, as the slower reading rates in textual paragraphs indicate, though this hypothesis on the different cognitive load of each kind of task should be tested in more detail in future research. However, dividing a broad issue into smaller questions and asking students to answer these questions, which seems a less demanding task, makes students concentrate on processing all of the information independent of its relevance for the task and causes more single-unit processing, as the focus is on searching specific information to answer specific questions. These different patterns of processing are coincident with those found by other researchers (Cerdán et al., 2007; Rouet et al., 2001).

Thus, integrating information from multiple documents in order to form a coherent representation from a subset of information in the documents requires the reader to concentrate on that information and process it more carefully than the rest of the information, which is similar to the behavior of good learners reported by Goldman (2004). It also means that when the reader finds information connected to other information he or she goes from one to the other in order to integrate the two pieces of information. This behavior is similar to the reinstatement and integration behavior reported by Mannes and Hoyer (1996). These two behaviors, which characterized the intertext essay task, are in opposition to the single-unit processing behavior fostered by asking students to answer more narrow intratext questions covering the same information.

Correlational data support these conclusions and the distinction between performance and learning (Schmidt & Bjork, 1992), as well as that between memory and learning (Kintsch, 1998). Regarding the first distinction, simply performing the task of either answering intratext questions or writing an essay with information explicit in a number of documents while having them available can be achieved by processing isolated units of information at a very superficial level, that is, searching, finding, and copying explicit information. In fact, performance on both tasks correlates neither with understanding and recalling, as measured by the sentence verification test, nor with deep comprehension, as measured by the transfer test. However, deep learning, as measured by the transfer test, depends mainly on integrating information across documents. With regard to the distinction between memory and learning, our data show that whereas scores on the sentence verification test, which involves mainly memory, are not related to integrating information, scores on the transfer test, which measures learning (Kintsch, 1998), are significantly related to integration. Thus, we provide evidence for the distinctions mentioned already in a situation not previously investigated, that is, the integration of infor-

mation from multiple documents, and we do so by using new processing measures.

Regarding the use of the think-aloud procedure in multiple text design, data from verbal protocols provide convergent evidence for behavioral results showing that the intertext essay task promoted a deeper processing of the multiple sources (i.e., more verbalizations about relevant units of information, which indicates deep comprehension, and fewer superficial writing verbalizations) in comparison to intratext questions. Our study also shows that making students think aloud globally increases the time needed to read the texts and perform the experimental task, as well as apparently interferes with deep learning results. Thinking aloud also made students read in a more homogeneous pattern without discriminating between relevant and irrelevant information, in comparison to the silent condition. Finally, thinking aloud tended to increase performance and sentence-verification scores, especially for students answering intratext questions.

Hence, as predicted by Ericsson and Simon (1980, 1993) thinking aloud created an increase in experimental time and a decrease in reading speed, a consequence of making students perform two tasks simultaneously (reading and thinking aloud). However, thinking aloud tended to increase text-based scores and decrease deep learning. Other authors, as reviewed earlier, also found interference effects for the think-aloud methodology. Wade and Trathen (1989) demonstrated that thinking aloud was detrimental for a task requiring the noting of ideas. Magliano et al. (1999) found that thinking aloud limited the higher level effects of induced reading strategies but facilitated memory for text. It seems that, at least for college students, thinking aloud forces an explicit concentration on surface textual information and limits the possibility of deeper processing activities because of cognitive overload (Sweller, 1994; Sweller, Van Merriënboer, & Paas, 1998). It may be that thinking aloud is regarded as an important assessment component in the experiment, hence students become committed to the think-aloud instructions, understanding them as instructions to paraphrase as much as possible and omit other deeper processing activities when reading the texts.

In sum, in the present experiment we confirmed that selecting relevant units of information and combining them by going from one to another are important processes for achieving integration from multiple documents, whereas the processing of isolated units of information seriously interferes with integration. However, integration across documents is highly influenced by the task that students are asked to perform in order to integrate information. The more integration a task requires, the better the results in terms of the learning it produces, although no differences in terms of a more superficial level of understanding are observed. It should be noted that the type of integration we dealt with in this study—that is, reading to select, process, and integrate a subset of information from multiple documents to describe and explain a physical phenomenon—is an important and complex skill in the knowledge society. Nowadays students are expected not only to understand a single text, but also to be able to integrate information from multiple sources and to use it for different purposes. We have just analyzed the role of different tasks in achieving such integration for university students. Other issues such as the role of sources' reliability, individual and developmental differences, the impact of previous background knowledge, and teaching strategies to foster this sort of integration still need to be investigated.

## References

- Britt, M. A., Rouet, J.-F., Georgi, M. C., & Perfetti, C. A. (1994). Learning from history texts: From causal analysis to argument models. In G. Leinhardt, I. L. Beck, & C. Stainton (Eds.), *Teaching and learning in history* (pp. 47–84). Hillsdale, NJ: Erlbaum.
- Cerdán, R., Vidal-Abarca, E., Martínez, T., Gilabert, R., & Gil, L. (2007). *Impact of question-answering tasks on search processes, text recall and comprehension*. Manuscript submitted for publication.
- Coté, N., & Goldman, S. R. (1999). Building representations of informational text: Evidence from children's think-aloud protocols. In H. Van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 169–193). Mahwah, NJ: Erlbaum.
- Coté, N., Goldman, S. R., & Saul, E. U. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processes*, 25, 1–53.
- Crain-Thoreson, C., Lippman, M. Z., & McClendon-Magnuson, D. (1997). Windows on comprehension: Reading comprehension processes as revealed by two think-aloud procedures. *Journal of Educational Psychology*, 89, 579–591.
- Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87, 215–251.
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.
- Fletcher, C. R. (1986). Strategies for the allocation of short-term memory during comprehension. *Journal of Memory and Language*, 25, 43–58.
- Goldman, S. R. (1997). Learning from text: Reflections on the past and suggestions for the future. *Discourse Processes*, 23, 357–398.
- Goldman, S. R. (2004, September). *Literacy in a knowledge society: Implications for comprehension research and practice*. Keynote address at the Text and Picture EARLI SIG meeting, Valencia, Spain.
- Goldman, S. R., & Durán, R. P. (1988). Answering questions from oceanography texts: Learner, task and text characteristics. *Discourse Processes*, 11, 373–412.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, England: Cambridge University Press.
- Long, D. L., & Bourg, T. (1996). Thinking aloud: Telling a story about a story. *Discourse Processes*, 21, 329–339.
- Magliano, J. P., & Graesser, A. C. (1991). A three-pronged method for studying inference generation in literary text. *Poetics*, 20, 193–232.
- Magliano, J. P., & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure and latent semantic analysis. *Discourse Processes*, 21, 251–283.
- Magliano, J. P., Trabasso, T., & Graesser, A. C. (1999). Strategic processing during comprehension. *Journal of Educational Psychology*, 91, 615–629.
- Mannes, S. (1994). Strategic processing of text. *Journal of Educational Psychology*, 86, 577–588.
- Mannes, S., & Hoyer, S. M. (1996). Reinstating knowledge during reading: A strategic process. *Discourse Processes*, 21, 105–130.
- Mannes, S. M., & Kintsch, W. (1987). Knowledge organization and text organization. *Cognition and Instruction*, 4, 91–115.
- McNamara, D. S., Kintsch, E., Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, 14, 1–43.
- McNamara, D. S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, 22, 247–288.
- Perfetti, C. A., Rouet, J.-F., & Britt, M. A. (1999). Toward a theory of document representation. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 99–123). Mahwah, NJ: Erlbaum.
- Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Erlbaum.
- Ramos, J. L., & Cuetos, F. (2000). *PROLEC-SE, evaluación de los pro-*



cesos lectores en alumnos del tercer ciclo de Educación Primaria y secundaria [Reading processes assessment for high-school students]. Madrid, Spain: Tea Ediciones.

Rouet, J.-F., Britt, M. A., Mason, R. A., & Perfetti, C. A. (1996). Using multiple sources of evidence to reason about history. *Journal of Educational Psychology, 88*, 478–493.

Rouet, J.-F., Favart, M., Britt, M. A., & Perfetti, C. A. (1997). Studying and using multiple documents in history: Effects of discipline expertise. *Cognition and Instruction, 15*, 85–106.

Rouet, J.-F., Vidal-Abarca, E., Bert-Erboul, A., & Millogo, V. (2001). Effects of information search tasks on the comprehension of instructional text. *Discourse Processes, 31*, 163–186.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*, 207–217.

Suh, S., & Trabasso, T. (1993). Inferences during reading: Converging evidence from discourse analysis, talk-aloud protocols and recognition priming. *Journal of Memory and Language, 32*, 279–300.

Sweller, J. (1994). Cognitive load theory, learning difficulty and instructional design. *Learning and Instruction, 4*, 295–312.

Sweller, J., Van Merriënboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*, 251–296.

Trabasso, T., & Magliano, J. P. (1996). How do children understand what they read and what can we do to help them? In M. Graves, P. Van den

Broek, & B. Taylor (Eds.), *The first R: A right of all children* (pp. 160–188). New York: Columbia University Press.

Trabasso, T., & Suh, S. (1993). Understanding text: Achieving explanatory coherence through on-line inferences and mental operations in working memory. *Discourse Processes, 16*, 3–34.

Vidal-Abarca, E., & Martínez, T. (2002). Read&Answer: A software to study on-line text learning and comprehension processes. Comprehension and learning from text research group, University of Valencia, Valencia, Spain.

Wade, S. E., & Trathen, W. (1989). Effect of self-selected study methods on learning. *Journal of Educational Psychology, 81*, 40–47.

Whitney, P., & Budd, D. (1996). Thinking-aloud protocols and the study of comprehension. *Discourse Processes, 21*, 341–351.

Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology, 91*, 301–311.

Wineburg, S. S. (1991). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology, 83*, 73–87.

Wineburg, S. S. (1994). The cognitive representation of historical texts. In G. Leinhardt, I. Beck, & C. Stainton (Eds.), *Teaching and learning in history* (pp. 85–135). Hillsdale, NJ: Erlbaum.

Wolfe, M. B. W., & Goldman, S. R. (2005). Relations between adolescents' text processing and reasoning. *Cognition and Instruction, 23*, 467–502.

Appendix

Intertext and Intratext Questions With Corresponding Text Ideas and Distributions Across Texts

QUESTIONS: CORRESPONDING IDEAS AND TEXT PARAGRAPHS		IDEA	TEXT
INTERTEXT	Explain how bacteria resist the effects of antibiotics and which biological mechanisms explain this phenomenon and its transmission to other bacteria.	22 IDEAS	P, G, R
INTRATEXT 1	Which characteristics of bacteria influence the development of bacterial resistance to antibiotics?	1, 2, 3, 4	R 2–1
INTRATEXT 2	Which biological mechanisms permit bacteria to become resistant to antibiotics?	5–13	P 1–3 or G 2–1
INTRATEXT 3	Can resistance be transmitted to other bacteria? If so, under which circumstances?	14–16	P 1–3 or R 2–1
INTRATEXT 4	How can bacteria resist antibiotics?	17–22	G 1–3

Note. P = *New Perspectives on Bacteria Resistance*; G = *Genetics of Bacteria Resistance*; R = *Anti-Bacteria Resistance*.

Received May 18, 2006  
Revision received June 21, 2007  
Accepted July 20, 2007 ■

# A Comparison of Three Measures of Cognitive Load: Evidence for Separable Measures of Intrinsic, Extraneous, and Germane Load

Krista E. DeLeeuw and Richard E. Mayer  
University of California, Santa Barbara

Understanding how to measure cognitive load is a fundamental challenge for cognitive load theory. In 2 experiments, 155 college students (ages = 17 to 22; 49 men and 106 women) with low domain knowledge learned from a multimedia lesson on electric motors. At 8 points during learning, their cognitive load was measured via self-report scales (mental effort ratings) and response time to a secondary visual monitoring task, and they completed a difficulty rating scale at the end of the lesson. Correlations among the three measures were generally low. Analyses of variance indicated that the response time measure was most sensitive to manipulations of extraneous processing (created by adding redundant text), effort ratings were most sensitive to manipulations of intrinsic processing (created by sentence complexity), and difficulty ratings were most sensitive to indications of germane processing (reflected by transfer test performance). Results are consistent with a triarchic theory of cognitive load in which different aspects of cognitive load may be tapped by different measures of cognitive load.

**Keywords:** cognitive load, measurement, education, multimedia, learning

Suppose a student viewed a narrated animation explaining how an electric motor works, such as that partially shown in the left panel of Figure 1. The lesson lasts about 6 min and explains how electricity flows from a battery and crosses a magnetic field, which in turn creates a force that moves a wire loop.

A major challenge in designing multimedia lessons such as the electric motor lesson is to be sensitive to the learner's cognitive load during learning. In particular, the lesson should be designed so that the amount of cognitive processing required for learning at any one time does not exceed the learner's processing capacity (Mayer, 2001, 2005a; Mayer & Moreno, 2003; Sweller, 1999, 2005). However, researchers have not yet reached consensus on how to measure cognitive load during learning or even whether the various cognitive load measures are tapping the same construct (Brünken, Plass, & Luetner, 2003; Paas, Tuovinen, Tabbers, & van Gerven, 2003).

According to a triarchic theory of cognitive load based on cognitive load theory (Sweller, 1999, 2005) and the cognitive theory of multimedia learning (Mayer, 2001, 2005a; Mayer & Moreno, 2003), there are three kinds of cognitive processing during learning that can contribute to cognitive load: (a) *extraneous processing*, in which the learner engages in cognitive processing that does not support the learning objective (and that is increased by poor layout such as having printed words on a page and their corresponding graphics on another page); (b) *intrinsic (or essential) processing*, in which the learner engages in cognitive processing that is essential for comprehending the material (and that depends on the complexity of material, namely the number of interacting elements that must be kept in mind at any one time);

and (c) *germane (or generative) processing*, in which the learner engages in deep cognitive processing such as mentally organizing the material and relating it to prior knowledge (and that depends on the learner's motivation and prior knowledge, as well as prompts and support in the lesson). Table 1 gives examples of how each of these three facets of cognitive load can be manipulated within our multimedia learning situation involving a narrated animation about how an electric motor works.

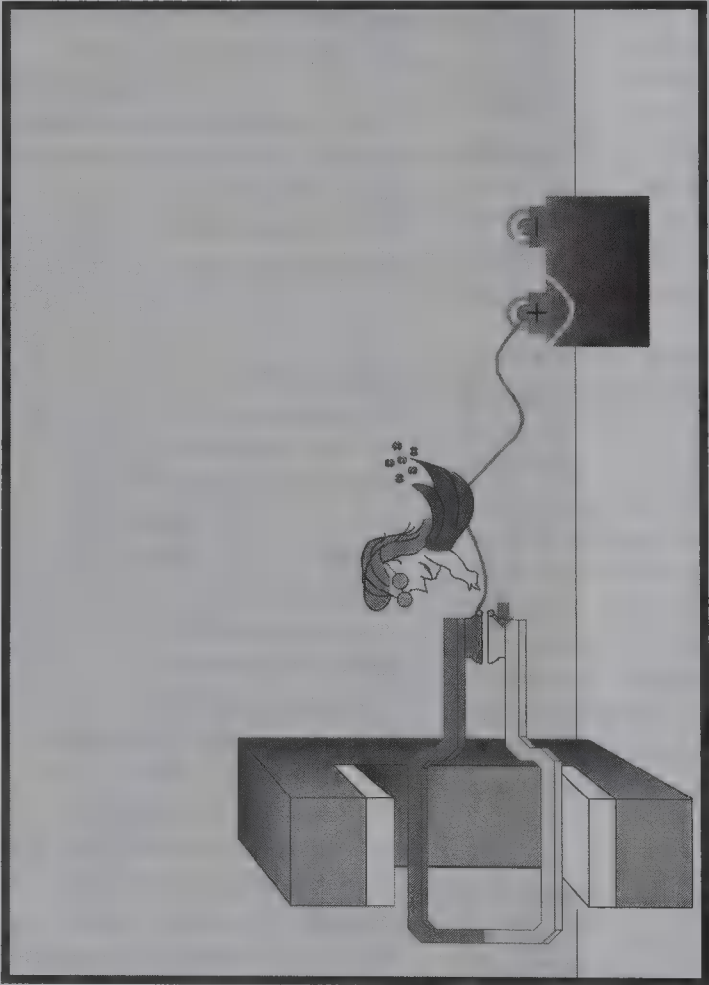
First, one way to manipulate extraneous processing in a multimedia lesson such as the electric motor lesson is through redundancy. A nonredundant lesson consists of presenting concurrent animation and narration, as summarized in the left panel of Figure 1; a redundant lesson consists of presenting concurrent animation, narration, and on-screen text, as summarized in the right panel of Figure 1. The on-screen text is redundant with the narration because both contain the same words and are presented at the same time. Redundancy can create extraneous cognitive load because the learner must expend precious cognitive capacity on reconciling the two verbal streams (i.e., checking to make sure the spoken words correspond to the printed words) and the learner must scan back and forth between the printed words and the animation (Mayer, 2005b). The processes of reconciling and scanning are forms of extraneous cognitive processing because they do not support the instructional objective (i.e., understanding how an electric motor works).

Second, one way to manipulate intrinsic processing in the electric motor lesson is through the complexity of the sentences. A low-complexity sentence (such as Sentence 1 in Figure 2) requires the learner to hold only a few concepts in working memory to understand the essential point, whereas a high-complexity sentence (such as Sentence 7 in Figure 2) requires the learner to hold many concepts in working memory to understand the essential point. Therefore, measures of cognitive load should reveal lower load after a low-complexity sentence than after a high-complexity sen-

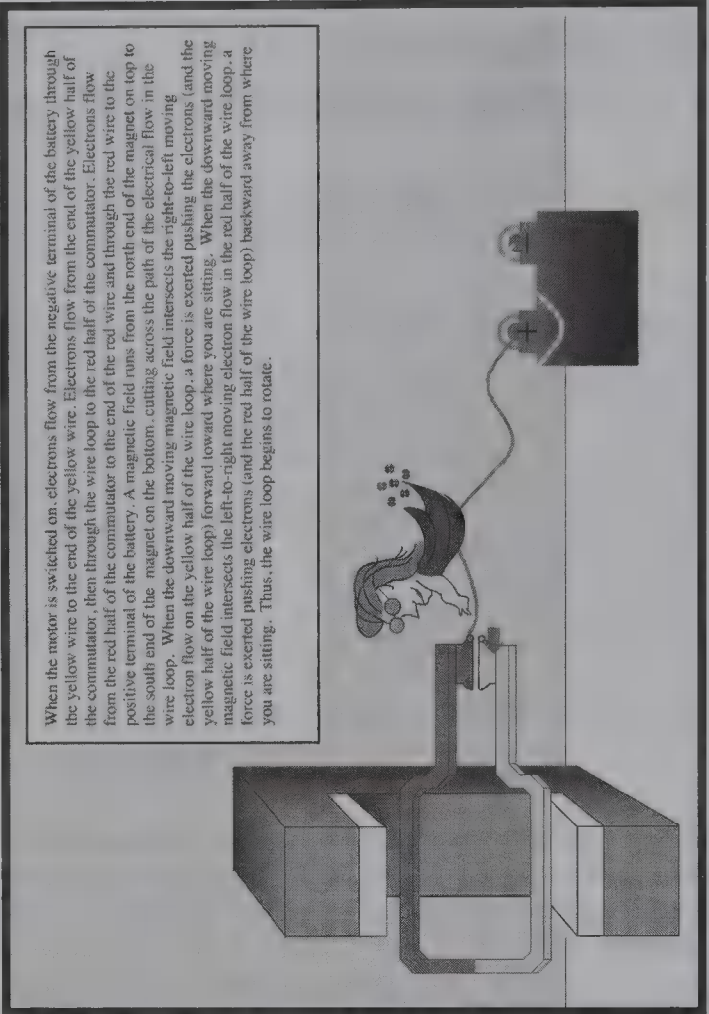
Correspondence concerning this article should be addressed to Krista E. DeLeeuw or Richard E. Mayer, Department of Psychology, University of California, Santa Barbara, CA 93106. E-mail: deleeuw@psych.ucsb.edu or mayer@psych.ucsb.edu



Non-Redundant



Redundant



Narration: “When the motor is switched on, electrons flow from the negative terminal of the battery through the yellow wire to the end of the yellow wire. Electrons flow from the end of the yellow wire through the commutator, then through the red half of the commutator, then through the wire loop to the red half of the commutator...”

Figure 1. Frames from nonredundant and redundant lessons on how an electric motor works.

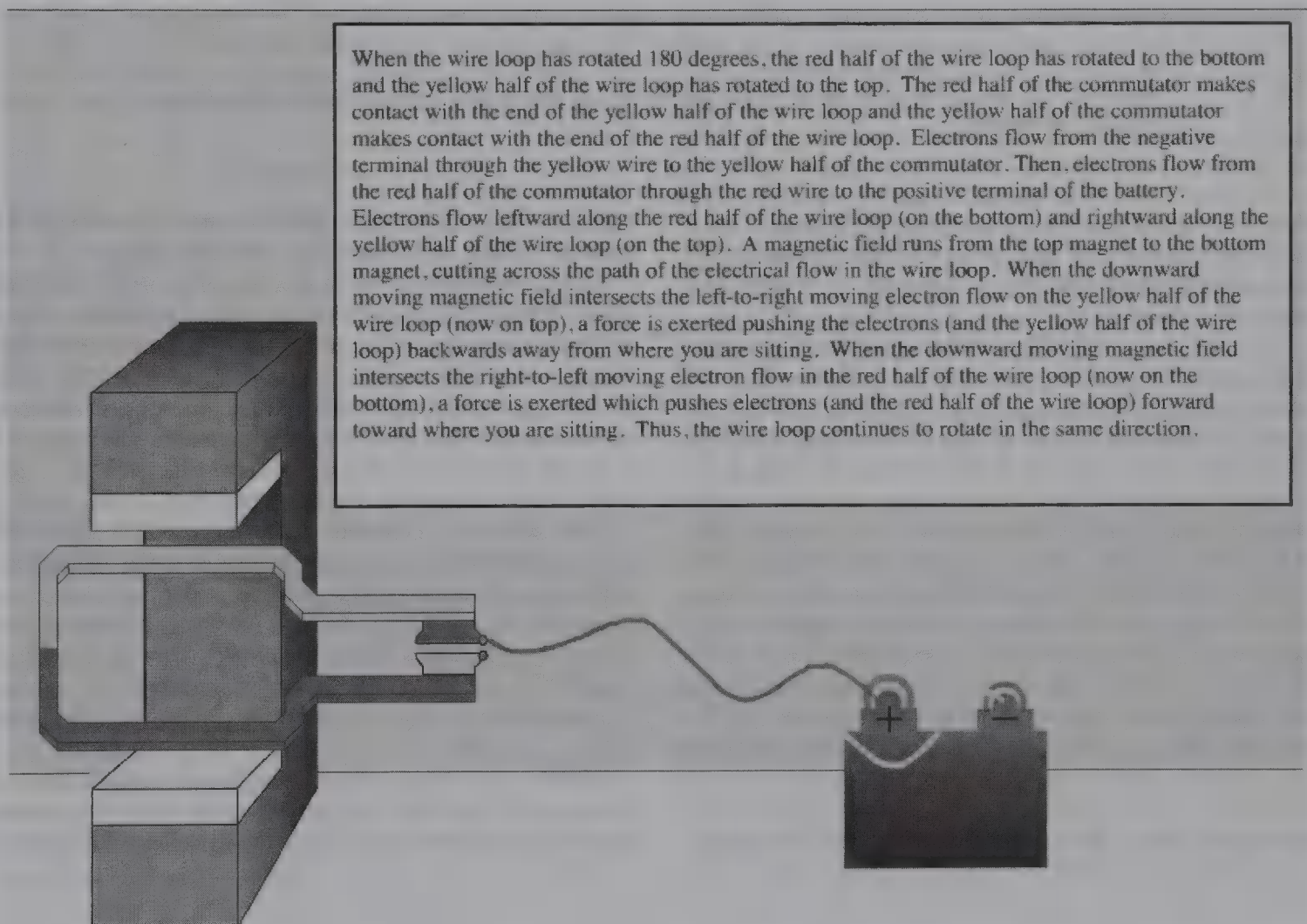
Table 1  
*Three Ways of Creating Cognitive Load in Multimedia Learning*

Type of cognitive load	Example of cognitive load manipulation
Extraneous load	Redundancy: Redundant presentation has concurrent animation, narration, and on-screen text; nonredundant presentation has concurrent animation and narration.
Intrinsic load	Complexity: A high-complexity sentence involves many interacting concepts; a low-complexity sentence involves few interacting concepts.
Germane load	Transfer: Low-transfer students are less likely to have engaged in germane processing during learning; high-transfer students are more likely to have engaged in germane processing during learning.

tence. Sentence complexity can create intrinsic cognitive processing because the learner needs to coordinate more pieces of information essential to comprehension (Mayer, 2005a; Sweller, 1999).

Third, one way to examine differences in germane processing in the electric motor lesson is to compare students who score

high on a subsequent test of problem-solving transfer with those who score low. High-transfer learners are more likely to have engaged in higher amounts of germane processing during learning than are low-transfer learners (Mayer, 2001, 2005a). Figure 3 shows a transfer test question along with answers from a



Low-complexity: Sentence 1. "When the loop has rotated 180 degrees, the red half of the wire loop has rotated to the bottom and the yellow half of the wire loop has rotated to the top."

High-complexity: Sentence 7. "When the downward moving magnetic field intersects with the left-to-right moving electron flow on the yellow half of the wire loop (now on top), a force is exerted pushing electrons (and the yellow half of the wire loop) backwards away from where you are sitting."

Figure 2. Examples of low-complexity and high-complexity sentences in a lesson on how an electric motor works.



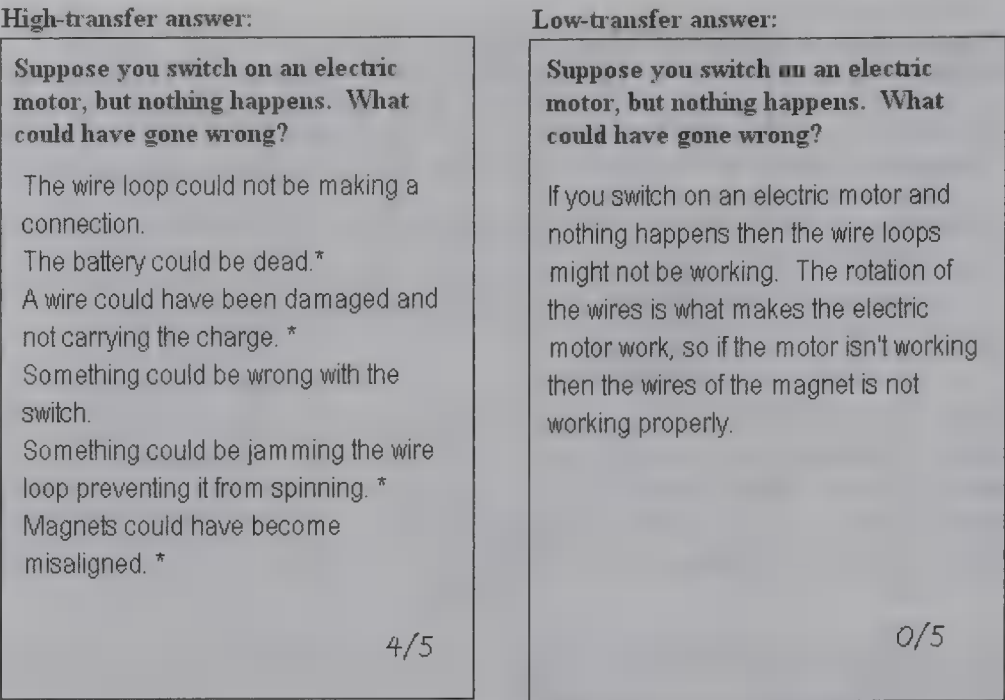


Figure 3. Examples of posttest answers from low-transfer and high-transfer students. Correct idea units are indicated by asterisks.

high-transfer student (in the left panel) and answers from a low-transfer student (in the right panel). Differences in transfer performance can be used to infer differences in germane processing during learning.

In this study, we examine the sensitivity of three commonly used techniques for measuring cognitive load—response time to a secondary task during learning, effort ratings during learning, and difficulty ratings after learning—to each of these three aspects of cognitive load. First, as shown in the first line of Table 2, we implemented a secondary task in our electric motor lesson. The secondary task was a visual monitoring task in which learners were asked to detect a periodic color change from pink to black in the background of the animation and to press the space bar as fast as possible each time this color change occurred. Response time has occasionally been used as a measure of cognitive load in multimedia research, with longer response times indicating greater cognitive load. When more cognitive resources are used by the primary task (learning from the narrated animation), fewer resources will be available to devote to the secondary task (noticing and responding to the background color change), resulting in a longer response time on the secondary task (Brünken, Steinbacher,

Schnotz, Plass, & Leutner, 2002; Chandler & Sweller, 1996; Marcus, Cooper, & Sweller, 1996; Sweller, 1988).

Second, as shown in the second line of Table 2, mental effort ratings were implemented in our electric motor lesson. We asked the learner to rate his or her current level of mental effort at points throughout the lesson on a 9-point scale ranging from 1 (*extremely low*) to 9 (*extremely high*). Self-reported mental effort ratings have been used somewhat more frequently in multimedia research than dual-task measurement of load (Paas et al., 2003; Paas & van Merriënboer, 1994; Paas, van Merriënboer, & Adam, 1994).

Third, as shown in the bottom line of Table 2, difficulty ratings were implemented in our electric motor lesson by asking the learner to make a retrospective judgment after the lesson concerning the lesson's difficulty, using a 9-point scale ranging from 1 (*extremely easy*) to 9 (*extremely difficult*). Self-reported difficulty ratings have been used somewhat less frequently in multimedia research (Kalyuga, Chandler, & Sweller, 1999; Mayer & Chandler, 2001).

A major goal of this study is to determine whether these three measures of cognitive load tap the same underlying construct (suggesting a unitary theory of cognitive load) or whether they are

Table 2  
Three Ways of Measuring Cognitive Load in Multimedia Learning

Type of measure	Implementation of measure
Response time to secondary task	At each of eight points in an animated narration, the background color slowly changes from pink to black, and the learner's task is to press the spacebar as soon as the background color changes.
Effort rating	At each of eight points in an animated narration the learner is asked to rate "your level of mental effort on this part of the lesson" on a 9-point scale ranging from 1 ( <i>extremely low mental effort</i> ) to 9 ( <i>extremely high mental effort</i> ).
Difficulty rating	At the end of the lesson, the learner is asked to rate "how difficult this lesson was" on a 9-point scale ranging from 1 ( <i>extremely easy</i> ) to 9 ( <i>extremely difficult</i> ).

sensitive to different manipulations of cognitive load (suggesting a model of cognitive load that is more consistent with a triarchic theory). According to a unitary theory of cognitive load, each of the three measures of cognitive load should be sensitive to each of the three types of cognitive load (i.e., indicating higher load for redundant vs. nonredundant groups, high-complexity vs. low-complexity sentences, and low-transfer vs. high-transfer learners), and the three should be highly correlated with one another. In other words, they should indicate one unitary measurement of cognitive load. According to a triarchic theory of cognitive load, it is possible that different measures of cognitive load may be sensitive to different types of cognitive load. If cognitive load is not a unitary construct and the measures are not all sensitive to the same type of load, then we should find that the three measures of cognitive load are not highly correlated with one another. If this is the case, the measures should be evaluated in terms of which aspect of the construct each measure best taps into. Predictions of possible ways in which the measures may relate to the three aspects of cognitive load in a triarchic theory are listed in Table 3. Alternatively, the possibility that the measures may gauge altogether different underlying constructs should be considered. We examined these questions in two experiments.

### Experiment 1

In Experiment 1, we tested the predictions of the unitary theory of cognitive load and the triarchic theory of cognitive load by examining the sensitivity of each of three common metrics of cognitive load—response time to a secondary task, mental effort rating, and difficulty rating—to manipulations intended to create differences in extraneous, intrinsic, and germane processing.

### Method

#### Participants and Design

The participants were 56 college students. There were 16 men and 40 women, and their ages ranged from 18 to 22. We used a  $2 \times 2$  mixed design with the between-subjects factor being redundancy (redundant vs. nonredundant lesson) and the within-subject factor being complexity (high- vs. low-complexity sentences within the lesson). Concerning redundancy, 28 participants served in the nonredundant group (in which they received a computer-based lesson containing animation and narration), and 28 participants served in the redundant group (in which they received the same computer-based lesson containing the animation and narration plus on-screen text). Concerning complexity, four of the sen-

tences in the lesson were identified as particularly high-complexity sentences (because they contained many interacting concepts), and four of the sentences in the lesson were identified as particularly low-complexity sentences (because they contained few interacting concepts). For each participant, cognitive load measurements were taken after each of the eight target sentences (for response time and effort rating) and after the entire lesson (for difficulty rating).

#### Materials and Apparatus

The computer-based materials consisted of a pretest questionnaire, a mental effort rating questionnaire, and two computer-based multimedia lessons on how an electric motor works. A Flash program designed for this experiment presented the multimedia lessons and collected data on the pretest questionnaire, response time to secondary tasks, and mental effort ratings. The paper-based materials consisted of a difficulty rating sheet and seven posttest problem sheets.

*Pretest questionnaire.* The computer-based pretest questionnaire assessed participants' previous experience with and knowledge about electronics and asked them to report their age, gender, SAT scores, and year in college.

*Multimedia lesson.* The nonredundant program (exemplified in the left portion of Figure 1) presented a narrated animation about how electric motors work that was used by Mayer, Dow, and Mayer (2003). The redundant version was identical except that it also presented on-screen text in paragraph form that was identical to the words spoken in the narration (as exemplified in the right panel of Figure 1). The on-screen text was redundant with the spoken narration, which is thought to increase extraneous cognitive load, reflected in lower scores on tests of transfer (Mayer, 2005b). Both versions of the lesson also included two levels of sentence complexity. We analyzed each sentence or clause in the script in terms of the number of interacting elements, or the number of idea units that must be kept in mind at one time for comprehension to take place. We then chose the four points of the lesson with the highest number of interacting elements, which we call *high sentence complexity*, and the four points with the lowest number of interacting elements, which we call *low sentence complexity*. We used these eight points during the lesson to implement the secondary task and the mental effort question.

The lesson was approximately 6 min long and described the cause-and-effect steps in the operation of an electric motor, which consisted of a battery, electrical wires, a commutator, a wire loop, and a pair of magnets.

*Secondary task and mental effort rating.* At eight points during the animation determined by the analysis of sentence complex-

Table 3  
Possible Relations Among Three Measures of Cognitive Load and Three Types of Cognitive Load

Type of measure	Type of cognitive load
Response time	Extraneous—When material includes redundant words, the learner wastes cognitive capacity, resulting in slower response times to a secondary task.
Effort rating	Intrinsic—When material is more complex, the learner must work harder to comprehend it, resulting in higher effort ratings.
Difficulty rating	Germane—Learners who perform well on the transfer test had capacity available for deeper processing during learning, reflected in lower difficulty ratings.



ity, the background color of the animation gradually changed from pink to black. Participants were instructed to press the space bar as quickly as they could when they saw the color begin to change. Four of the color changes occurred at the end of high-complexity sentences and four occurred at the end of low-complexity sentences. The program recorded the students' response times (RTs; in milliseconds) to press the space bar for each of the eight events and calculated an average RT for high-complexity sentences and low-complexity sentences for each student. When the student pressed the space bar, the program paused the animation and presented a question regarding the amount of mental effort the participant was currently experiencing. The question asked participants to "please rate your level of mental effort on this part of the lesson" and gave them a Likert-type scale ranging from 1 (*extremely low mental effort*) to 9 (*extremely high mental effort*) from which to choose their response. The program computed and recorded the mean effort rating for each student for the four high-complexity sentences and the four low-complexity sentences. After indicating their answer by clicking the corresponding rating, participants had a choice of two buttons to press: continue or replay. Although participants believed that these buttons allowed them to either replay the last section of the animation or to continue from where they left off, the buttons actually performed identical functions; they both replayed the last sentence and continued onward. All students opted for the continue button.

**Difficulty rating and posttest.** The difficulty rating scale consisted of a sheet of paper containing the instruction "Please indicate how difficult this lesson was by checking the appropriate answer"; response choices ranged from 1 (*extremely easy*) to 9 (*extremely difficult*). The posttest consisted of seven sheets of paper, each with a question printed at the top and space for participants to write their answers. The questions tested participants' ability to transfer information they had learned about electric motors to problem-solving situations and were identical to those used by Mayer et al. (2003). The questions were (a) "What could you do to increase the speed of the electric motor, that is, to make the wire loop rotate more rapidly?" (b) "What could you do to increase the reliability of the electric motor, that is, to make sure it would not break down?" (c) "Suppose you switch on an electric motor, but nothing happens. What could have gone wrong?" (d) "What could you do to reverse the movement of the electric motor, that is, to make the wire loop rotate in the opposite direction?" (e) "Why does the wire loop move?" (f) "If there was no momentum, how far would the wire loop rotate when the motor is switched on?" and (g) "What happens if you move the magnets further apart? What happens if you connect a larger battery to the wires? What happens if you connect the negative terminal to the red wire and the positive terminal to the yellow wire?"

**Apparatus.** The apparatus consisted of five Sony Vaio laptop computers with 15-in. screens and Panasonic headphones.

### Procedure

Participants were randomly assigned to the redundant or nonredundant group and were seated in a cubicle at a work desk with an individual laptop computer. There were between 1 and 5 participants in each session. Figure 4 summarizes the procedure. First, the experimenter briefly described the study, and each participant read and signed an informed consent form, which was collected by

the experimenter. Second, the experimenter logged the participants in on their respective computers and asked them to complete the pretest questionnaire presented on the computer screen. Next, participants read on-screen instructions about how to perform the secondary task, and then they completed a practice task. The practice task was identical to the experimental dual task, except that it presented an animation on an unrelated topic (i.e., how brakes work). After the practice task, the experimenter inquired as to whether all participants were comfortable with the task, and participants were given a chance to ask questions. Once all participants indicated that they fully understood the task, they continued on to the experimental phase of the experiment in which the electric motor lesson was presented. Participants in the redundant group received the redundant version and participants in the nonredundant group received the nonredundant version. After the lesson, participants were given the difficulty rating sheet to complete at their own pace, followed by each of the seven posttest sheets. The posttest questions were presented one at a time in the order indicated in the *Materials and Apparatus* section, and participants were given a limit of 3 min to write their answers to each question.

## Results and Discussion

### Scoring

RT measurements more than 3 standard deviations from the mean of any of the RT measurements were replaced with that participant's series mean, with respect to whether the outlying RT occurred at a high- or low-complexity sentence. We opted to replace the outlying RT rather than simply exclude each outlying RT or the entire participant, with the assumptions that a RT of more than 3 standard deviations above the mean indicated the participant's inattention and that averaging across that participant's low- or high-complexity trials (depending on the type of trial in which the outlier occurred) would provide a sufficient estimate of how that participant would have performed on that trial had he or she given the task the appropriate attention. In cases in which more than one RT was an outlier within a participant's series, that participant was excluded from further analysis because there was not enough information to make a sufficient estimate of RT on those trials. This resulted in 2 participants being excluded, leaving 26 in the redundant group and 28 in the nonredundant group ( $N = 54$ ).

Answers on the transfer test were coded for correct answers, out of a total possible 25. The list of acceptable answers was generated by Mayer et al. (2003). For example, the acceptable answers for the third question about what went wrong were "the wire loop is stuck," "the wire is severed or disconnected from the battery," "the battery fails to produce voltage," "the magnetic field does not intersect the wire loop," and "the wire loop does not make contact with the commutator." Students did not receive credit for vague answers, such as "Something is wrong with the battery," but did get credit for correct answers that were worded differently than in the lesson. For all but one question, there was more than one possible acceptable answer. Each acceptable answer generated by the participant earned 1 point; we summed the number of points across all of the questions to obtain a total transfer score, resulting in a range from 0 to 16 correct idea units, with a median of 6.5.

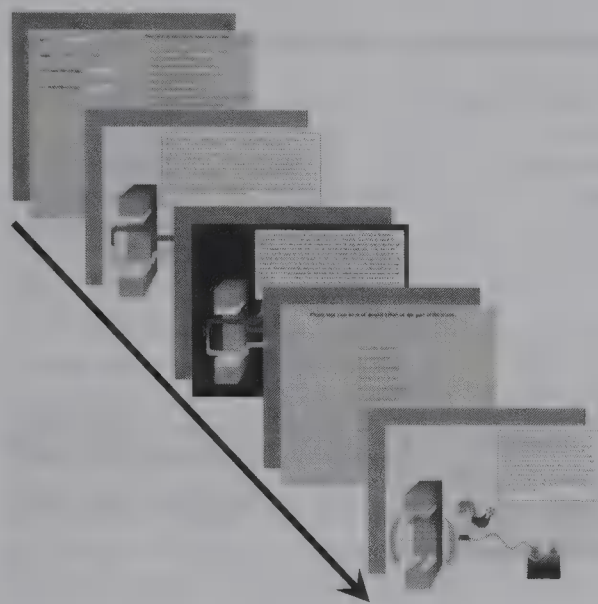
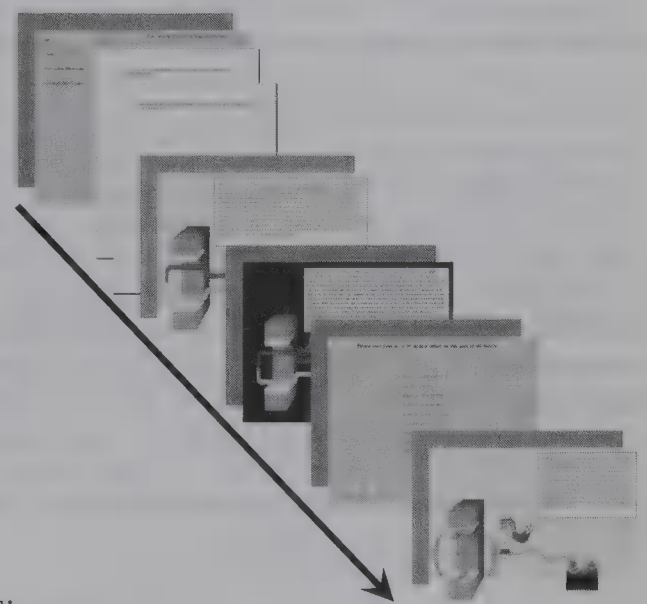
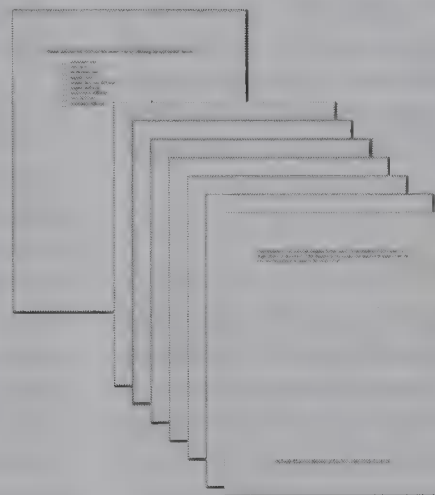
Experiment 1Experiment 2Both Experiments:

Figure 4. Summary of procedure.

Assignment to low- versus high-transfer groups was based on a median split in which students who scored 7 points or more were assigned to the high-transfer group ( $n = 27$ ) and students who scored 6 points or less were assigned to the low-transfer group ( $n = 27$ ).

*Response Time to the Secondary Task*

If response time to a secondary task during learning is a valid measure of cognitive load, it should differ on the basis of the manipulations thought to affect cognitive load. That is, response time should be longer for the redundant group than for the nonredundant group, on more complex sentences in the lesson than on less complex sentences, and for low-transfer learners than for high-transfer learners.

The first row of data in Table 4 shows the mean response times (and standard deviations), first for the redundant and nonredundant groups, second for the high- and low-complexity sentences, and finally for the low- and high-transfer learners. A box around two means (and their standard deviations) indicates that we found some difference between them, with bold lines indicating a significant

difference at  $p < .05$  and light lines indicating a nonsignificant difference (at  $p < .10$ ) but an effect size greater than .20. We conducted a 2 (redundant vs. nonredundant)  $\times$  2 (low- vs. high-complexity sentences) mixed analysis of variance (ANOVA), with redundancy as a between-subjects factor and complexity as a repeated-measures factor. First, there was a marginal main effect for redundancy, with the redundant group producing marginally higher response times to the secondary task than the nonredundant group,  $F(1, 52) = 3.77, p = .06, d = .53$ . Although this trend was not statistically significant, the effect size is in the medium range, which suggests that the lack of significance may have been the result of a small sample size (Cohen, 1988). Second, there was a marginal main effect for complexity, with high-complexity sentences requiring marginally longer RTs to the secondary task than did low-complexity sentences,  $F(1, 52) = 3.85, p = .06, d = .25$ . Again, we note that although this trend was not statistically significant, the effect size was in the small range. No significant interaction was found between complexity and redundancy,  $F(1, 52) = 0.004, ns$ . Third, a  $t$  test revealed that students who scored low on transfer performance did not differ significantly on re-



Table 4  
Means (and Standard Deviations) for Three Types of Cognitive Load Manipulations Based on Three Measures of Cognitive Load:  
Experiment 1

Measure of cognitive load	Type of cognitive load					
	Extraneous load: Redundancy (Which cognitive load measure(s) is sensitive to redundancy?)		Intrinsic load: Complexity (Which cognitive load measure(s) is sensitive to sentence complexity?)		Germane load: Transfer (Which cognitive load measure(s) is sensitive to transfer performance?)	
	Redundant	Nonredundant	High	Low	High ( <i>n</i> = 27)	Low ( <i>n</i> = 27)
Response time (ms)	2,657 (825)	2,249 (719)	2,555 (1,035)	2,337 (714)	2,477 (933)	2,414 (636)
Effort rating	4.97 (1.54)	5.49 (1.41)	5.43 (1.55)	5.05 (1.50)	5.30 (1.57)	5.18 (1.42)
Difficulty rating	5.15 (1.54)	5.36 (1.47)			4.63 (1.39)	5.89 (1.34)

Note. *N* = 54. Boxes with bold lines indicate significant difference ( $p < .05$ ). Boxes with light lines indicate  $.05 < p < .10$  and effect size greater than  $d = .20$ .

sponse time compared with students who scored high on transfer performance,  $t(52) = -0.29$ , *ns*. We furthermore computed a Cronbach's alpha for the four RTs at low-complexity points in the lesson and for the four RTs at high-complexity points to investigate the internal reliability of this measurement. Although the low-complexity points showed a low reliability ( $\alpha = .33$ ), high-complexity points were shown to be reliable ( $\alpha = .70$ ).

#### Mental Effort Ratings During Learning

For self-reported mental effort during a lesson to be considered a valid measure of cognitive load, several conditions should be met. First, lessons containing animation with redundant text and narration should cause learners to rate their mental effort higher overall than nonredundant lessons containing only animation and narration. Second, learners should rate their mental effort significantly higher at difficult (high-complexity) points in the lesson than at simpler (low-complexity) points in the lesson. Third, students who score low on transfer should rate their mental effort significantly higher overall than those who score high on transfer (alternatively, higher effort may indicate more germane processing, which should lead to higher scores on the transfer test).

The second row of data in Table 4 shows the mean mental effort ratings (and standard deviations) for each of three comparisons, first for the redundant and nonredundant groups, second for the high- and low-complexity sentences, and last for the low- and high-transfer learners. We conducted a 2 (redundant vs. nonredundant)  $\times$  2 (high- vs. low-complexity sentences) mixed ANOVA, with redundancy as a between-subjects factor and complexity as a repeated-measures factor. First, there was no significant effect for redundancy, with the redundant group and the nonredundant group rating their mental effort equivalently,  $F(1, 52) = 1.67$ , *ns*. Second, in contrast, there was a significant main effect of complexity in which participants rated their mental effort as higher on high-complexity sentences than on low-complexity sentences,  $F(1, 52) = 17.78$ ,  $p < .001$ ,  $d = .25$ . No significant interaction was found between redundancy and complexity,  $F(1, 52) = 0.004$ , *ns*. Third, a *t* test revealed that students who scored low on transfer performance did not differ significantly on mental effort ratings than students who scored high on transfer performance,  $t(52) =$

$-0.30$ , *ns*. Cronbach's alpha indicated that measures of mental effort were reliable both at low-complexity points ( $\alpha = .84$ ) and at high-complexity points ( $\alpha = .90$ ).

#### Overall Difficulty Ratings After Learning

The last row of data in Table 4 shows the mean lesson difficulty ratings (and standard deviations) for each of two comparisons: between the redundant and nonredundant groups and between the low-transfer and high-transfer learners. There were no separate difficulty ratings for low and high complexity because this rating was only administered at the end of the lesson. If the difficulty rating is a valid measure of cognitive load, then the mean difficulty ratings should differ between redundant and nonredundant versions of that lesson. In addition, the mean difficulty rating of the low-transfer group should be greater than the mean difficulty rating of the high-transfer group. Finally, although we would expect learners to rate high-complexity sentences as more difficult than low-complexity sentences, there was no way to examine this because difficulty ratings were solicited only for the entire lesson after learning.

First, a *t* test showed that ratings of difficulty did not differ between the redundant and nonredundant groups,  $t(52) = 0.50$ ,  $p = .62$ , which could indicate either that this rating scale is not sensitive to changes in extraneous load or that the redundant version of the multimedia lesson created no more cognitive load than the nonredundant version. The results of a second *t* test revealed that those who scored lower on the transfer test tended to rate the lesson as more difficult than those who scored higher,  $t(52) = 3.39$ ,  $p < .001$ ,  $d = .92$ .

#### Are Commonly Used Metrics of Cognitive Load Related to One Another?

If all of the methods we used to measure cognitive load did indeed measure the same construct, then we would expect to see significant correlations among all of the measures. Specifically, we would expect response time on a secondary task, self-reported mental effort, and difficulty rating to all be positively correlated. As shown in Table 5, however, we did not find this pattern of

Table 5  
Correlation Matrix of Dependent Measures for Both Groups:  
Experiment 1

Measure	1	2	3	4
1. Dual-task reaction time	—	.27*	.20	.07
2. Self-report mental effort		—	.26	.19
3. Lesson difficulty rating			—	-.22
4. Score on transfer test				—

Note.  $N = 54$ .

\*  $p < .05$ .

correlations. Of the six pairwise correlations, only one reached statistical significance; response time on the secondary task was significantly correlated with mental effort ( $r = .27$ ,  $p = .05$ ), but at a modest level. The correlation results do not support the contention that various cognitive load measures tend to measure the same thing, as proposed by the unitary theory of cognitive load.

#### Are Different Cognitive Load Measures Sensitive to Different Types of Cognitive Load?

Table 6 shows whether each cognitive load measure detected a significant difference for each type of cognitive load and the effect size indicated by each measure for each type of load. The pattern that emerges from these results suggests that each of the three measures—response time, effort rating, and difficulty rating—was sensitive mainly to one aspect of cognitive load. Response times were most sensitive to redundancy (which was intended to create extraneous cognitive processing), effort ratings during learning were most sensitive to complexity (which was intended to create intrinsic cognitive processing), and difficulty ratings after learning were most sensitive to transfer performance (which was intended to tap germane processing). Given the distinct pattern that was generated in Experiment 1, it is worthwhile to determine whether the pattern could be replicated in a similar study.

Overall, these commonly used measures of cognitive load do not appear to be related to one another in a clear-cut fashion. In contrast, the pattern of the results suggests that these measures may be tapping different aspects of cognitive load or different constructs altogether. The results are more consistent with the triarchic theory of cognitive load than with a unitary theory of cognitive load.

Table 6  
Effect Sizes for Three Types of Cognitive Load Manipulation Created Based on Three Measures of Cognitive Load

Measure of cognitive load	Type of cognitive load					
	Extraneous load (redundant vs. nonredundant)		Intrinsic load (high vs. low complexity)		Germane load (high vs. low transfer)	
	Exp. 1	Exp. 2	Exp. 1	Exp. 2	Exp. 1	Exp. 2
Response time	.53 <sup>†</sup>	.48*	.25 <sup>†</sup>	<i>ns</i>	<i>ns</i>	<i>ns</i>
Effort rating	<i>ns</i>	.35*	.25**	.16**	<i>ns</i>	<i>ns</i>
Difficulty rating	<i>ns</i>	<i>ns</i>			.92**	.48*

Note. Cohen's  $d$  is the measure of effect size. Difficulty rating does not apply to intrinsic load. Exp. = Experiment.

<sup>†</sup>  $p < .10$ . \*  $p < .05$ . \*\*  $p < .01$ .

## Experiment 2

In Experiment 1, an interesting pattern of results emerged in which each of the three measures of cognitive load tended to be sensitive to a different aspect of cognitive load. However, the many measures of cognitive load may have been somewhat intrusive to the learning task, resulting in measures that may reflect distraction rather than cognitive load per se. To encourage learners to focus on deep cognitive processing as their goal, in Experiment 2 we provided learners with *pretest questions* that we asked them to be able to answer at the end of the lesson. These pretest questions were shown by Mayer et al. (2003) to improve scores on the transfer test for this lesson. We reasoned that if students were more motivated to learn the information, they would pay closer attention to the lesson and therefore provide more valid measurements of cognitive load.

### Method

#### Participants and Design

Participants in Experiment 2 were 99 college students (33 male, 66 female) ranging in age from 17 to 22. Half of the participants were randomly assigned to the nonredundant group ( $n = 49$ ) and half to the redundant group ( $n = 50$ ). The design of Experiment 2 was identical to that of Experiment 1.

#### Materials, Apparatus, and Procedure

The materials, apparatus, and procedure of this experiment were identical to those of Experiment 1, with the addition of two pretest questions. The pretest questions provided the participants with knowledge about the type of information they were expected to learn during the lesson. They were presented on an 8.5- × 11-in. sheet before the lesson and consisted of Questions a and b from the transfer test (i.e., "What could you do to increase the speed of the electric motor, that is, to make the wire loop rotate more rapidly?" and "What could you do to increase the reliability of the electric motor, that is, to make sure it would not break down?"). Participants were told that the pretest questions were "representative of the types of questions that would be asked later." Participants were allowed to look over the pretest question sheet until they felt comfortable with the material, at which time they handed the sheet back to the experimenter. Then the experimenter reminded participants of the instructions, answered any questions, and instructed



them to begin the computer program. From this point on, the procedure was identical to Experiment 1.

Results

Scoring

Outlying RTs to the secondary task were dealt with in the same manner as in Experiment 1. Again, participants who had more than two outlying RTs in a series were excluded from further analyses. This resulted in the exclusion of only 3 participants, leaving 47 in the nonredundant group and 49 in the redundant group ( $N = 96$ ). Answers on the transfer test were coded in the same fashion as in Experiment 1.

Scores on the transfer test ranged from 0 to 13 correct idea units, with a median of 6.0, out of a possible total of 25. A subset of the data was coded by a second rater, resulting in a correlation of .772 ( $p = .009$ ). A median split resulted in 45 high scorers (scoring 7 points or more) and 51 low scorers (scoring 6 points or less).

RT to the Secondary Task

As in Experiment 1, we expected RT on the secondary task to be longer for the redundant group than for the nonredundant group, longer for high-complexity sentences in the lesson than for low-complexity sentences, and longer for low-transfer learners than for high-transfer learners.

The first row of data in Table 7 shows the mean RTs (and standard deviations) for three comparisons, first for the redundant and nonredundant groups, second for the high- and low-complexity sentences, and last for the low- and high-transfer learners. As in Table 4, a box around two means (and their standard deviations) indicates that we found some difference between them, with bold lines indicating a significant difference. We conducted a 2 (redundant vs. nonredundant)  $\times$  2 (low vs. high complexity) mixed ANOVA with redundancy as a between-subjects factor and complexity as a repeated measures factor. There was a significant main effect for redundancy, with the redundant group producing longer RTs than the nonredundant group,  $F(1, 94) = 5.44, p = .02, d = .48$ . However, no main effect of complexity was found,  $F(1, 94) = 1.67, ns$ , nor was there a significant interaction between trial type and redundancy,  $F(1,$

94) = 0.93, *ns*. These results show that redundant on-screen text caused longer RTs and indicate that participants who saw the nonredundant version had more free cognitive resources. However, unlike in Experiment 1, RT did not appear to be sensitive to the number of interacting elements (i.e., sentence complexity) at a given point in the lesson. Finally, a *t* test showed that there was no significant difference between low- and high-transfer learners on their RTs to the secondary task,  $t(94) = 1.68, ns$ . Cronbach's alpha showed internal reliability for RT measurements both at low-complexity points ( $\alpha = .79$ ) and at high-complexity points ( $\alpha = .76$ ).

Mental Effort Rating During Learning

As in Experiment 1, on the basis of unitary theory, we expected mental effort ratings to be higher (indicating more mental effort expended) for the students learning from the redundant lesson than for those learning from the nonredundant lesson. We also expected learners to rate their mental effort higher at difficult (high-complexity) sentences in the lesson than at easier (low-complexity) sentences in the lesson. Finally, we expected that learners who scored low on the transfer test would rate their mental effort higher overall than those who scored high on the transfer test.

The second row of data in Table 7 shows the mean RTs (and standard deviations) for three comparisons, first for the redundant and nonredundant groups, second for the high- and low-complexity sentences, and last for the low- and high-transfer learners. We conducted a 2 (redundant vs. nonredundant)  $\times$  2 (high vs. low complexity) mixed ANOVA with redundancy as a between-subjects factor and trial type as a repeated measures within-subjects factor. First, unlike in Experiment 1, ratings of mental effort differed significantly between the redundant and nonredundant lessons,  $F(1, 94) = 4.17, p = .04, d = .35$ . Second, we again found a significant main effect of complexity; learners rated their mental effort as higher on high-complexity sentences than on low-complexity sentences,  $F(1, 94) = 20.36, p < .001, d = .16$ . No significant interaction of complexity and redundancy was found,  $F(1, 94) = 0.21, ns$ . Third, a *t* test showed that learners who scored low on the transfer test had overall mental effort ratings similar to those who scored high on the transfer test,

Table 7  
Means (and Standard Deviations) for Three Types of Cognitive Load Manipulations Based on Three Measures of Cognitive Load: Experiment 2

Measure of cognitive load	Type of cognitive load					
	Extraneous load: Redundancy (Which cognitive load measure(s) is sensitive to redundancy?)		Intrinsic load: Complexity (Which cognitive load measure(s) is sensitive to sentence complexity?)		Germane load: Transfer (Which cognitive load measure(s) is sensitive to transfer performance?)	
	Redundant	Nonredundant	High	Low	High ( $n = 45$ )	Low ( $n = 51$ )
Response time (ms)	2,918 (872)	2,520 (797)	2,677 (869)	2,769 (974)	2,569 (848)	2,859 (847)
Effort rating	5.67 (1.59)	4.99 (1.67)	5.47 (1.66)	5.21 (1.68)	5.58 (1.58)	5.13 (1.71)
Difficulty rating	5.33 (1.83)	5.21 (1.74)			4.82 (1.92)	5.67 (1.56)

Note. Boxes with bold lines indicate significant difference ( $p < .05$ ).

$t(94) = -1.31$ , *ns*. Cronbach's alpha showed internal reliability for mental effort measurements both at low-complexity points ( $\alpha = .90$ ) and at high-complexity points ( $\alpha = .90$ ).

### Overall Difficulty Rating After Learning

As in Experiment 1, according to the unitary theory, we expected learners in the redundant group to rate the difficulty of the lesson as higher than those in the nonredundant group. In addition, we expected that learners who scored low on the transfer test would rate the lesson as more difficult than those who scored high.

The last row of data in Table 7 shows the mean lesson difficulty ratings (and standard deviations) for two comparisons—for the redundant and nonredundant groups and for the low- and high-transfer learners. There were no separate difficulty rating scores for low and high complexity because this rating was only administered at the end of the lesson. First, there was no significant difference between the redundant and nonredundant groups; learners in both groups rated the difficulty of the lesson similarly,  $t(94) = -0.31$ , *ns*. In contrast, there was a significant difference between high- and low-transfer learners; learners who scored low on the transfer test tended to rate the lesson as more difficult than those who scored high,  $t(94) = 2.38$ ,  $p = .02$ ,  $d = .48$ .

### Are Commonly Used Metrics of Cognitive Load Related to One Another?

Table 8 shows the correlations among the dependent variables. If these methods are all measuring a unitary or overall level of cognitive load, we would expect them all to be significantly positively correlated with one another. However, RT to the secondary task was not positively correlated with any of the other measures of cognitive load. It was, however, significantly negatively correlated with scores on the transfer test ( $r = -.30$ ,  $p = .003$ ), indicating that participants who scored lower on the transfer test tended to take longer to respond to the secondary task, which is the result we would expect if cognitive load were causing a slower RT. Mental effort ratings were positively correlated with difficulty ratings ( $r = .33$ ,  $p = .001$ ) but with no other measure. Difficulty ratings were negatively correlated with transfer scores ( $r = -.22$ ,  $p = .03$ ) but with no other measure. As can be seen by comparing Tables 5 and 8, the measures were more correlated with one another in Experiment 2 than in Experiment 1, but most of the relations were weak or nonsignificant.

Table 8  
Correlations Between Dependent Measures for Both Groups:  
Experiment 2

Measure	1	2	3	4
1. Dual-task reaction time	—	.12	.13	-.30**
2. Self-reported mental effort		—	.33**	.11
3. Lesson difficulty rating			—	-.22*
4. Score on transfer test				—

Note.  $N = 96$ .  
\*  $p < .05$ . \*\*  $p < .01$ .

### Are Different Cognitive Load Measures Sensitive to Different Types of Cognitive Load?

Table 6 shows whether each cognitive load measure detected a significant difference for each type of cognitive load and the effect size indicated by each cognitive load measure for each type of cognitive load. The pattern of results for Experiment 2 is quite similar to that in Experiment 1.

In Experiment 2, it was again apparent that RTs on the secondary task were most sensitive to redundancy than to other manipulations of cognitive load. Because redundancy was intended to create extraneous cognitive load, we can conclude that RTs were most sensitive to differences in extraneous load. Ratings of mental effort were most sensitive to the complexity of the sentences, which was intended to create intrinsic cognitive load. Finally, ratings of the overall difficulty of the lesson were most sensitive to differences in the learning outcomes of the students, in terms of their scores on the transfer test, which is an indication of germane cognitive load. Taken together with the findings from Experiment 1, these results show that these measures are tapping separable aspects of cognitive load.

### Conclusion

#### Summary of Results

Across two experiments, we found that different measures of cognitive load were sensitive to different types of cognitive load: (a) RT to the secondary task was most sensitive to manipulations of extraneous processing (reflected in longer RTs for the redundant group than for the nonredundant group), (b) effort ratings during learning were most sensitive to manipulations of intrinsic processing (reflected in higher effort attributed to high-complexity sentences than to low-complexity sentences), and (c) difficulty ratings after learning were most sensitive to differences related to germane processing (reflected in higher difficulty reported by low-transfer learners than by high-transfer learners). In both experiments, we also found that different measures of cognitive load were not highly correlated with one another. This work is consistent with recent findings reported by Ayres (2006), in which a subjective rating measure administered after each problem was sensitive to differences in intrinsic processing, and by Brünken, Plass, and Leutner (2004), in which a dual-task measure was sensitive to differences in extraneous processing.

#### Theoretical Implications

If cognitive load is a unitary construct, reflecting an overall amount of cognitive resources allocated to the task, all types of manipulations of the learning situation (be it a manipulation of the study materials or of the motivation of the learner, etc.) should cause a corresponding change in the amount of load, and all measures that are directly related to cognitive load should correlate with one another. However, if cognitive load is composed of, or influenced by, different elements, as proposed by Sweller (1999) and Mayer (2001), then different manipulations of the learning situation can cause different types of cognitive load to vary in distinguishable ways. In this case, it may be possible that some measures are more sensitive to one type of change in cognitive load than to others.



On the theoretical side, this pattern of results is most consistent with a triarchic theory of cognitive load, in which cognitive processing during learning is analyzed into three capacity-demanding components—extraneous processing, intrinsic (or essential) processing, and germane (or generative) processing. More important, this study provides empirical support, replicated across two experiments, for a dissociation among these three types of cognitive load. The results are not consistent with a unitary theory of cognitive load, which focuses on the overall amount of cognitive processing during learning.

### Practical Implications

On the practical side, the results provide some validation for each of the three measures of cognitive load—RT to a secondary task, mental effort rating during learning, and difficulty rating made retrospectively immediately following learning—because each measure was sensitive to a particular cognitive load manipulation. An important practical implication is that different measures of cognitive load should not be assumed to measure overall cognitive load, but may be effectively used to measure different types of cognitive load. In particular, when the goal is to assess the level of extraneous cognitive load, RT to a secondary task appears to be most appropriate; when the goal is to assess the level of intrinsic cognitive load, mental effort ratings during learning may be most appropriate; and when the goal is to detect the learner's level of germane cognitive load, a simple difficulty rating immediately after learning may prove most useful.

### Limitations and Future Directions

Although a similar pattern of results was obtained across two experiments, there is still a need for replication studies involving different instructional materials and different learners. It is possible that given different learning materials or different learner characteristics, we may find a different pattern of results. Regarding the intercorrelation of the measures in the two experiments, it is possible that we found mostly small, nonsignificant effects in Experiment 1 because of a small sample size, but this argument does not hold for Experiment 2 because the sample size was nearly double.

Consistent with research on the expertise reversal effect (Kalyuga, 2005), we suspect that the pattern of results might depend on the prior knowledge of the learner. In our experiments, the learners generally had low prior knowledge, so we suggest that future research should also take the learner's level of prior knowledge into account.

In this study, we focused on three specific measures of cognitive load, but there are alternative ways to implement each measure and there are certainly other measures of cognitive load. Similarly, we examined three cognitive load manipulations, but alternative methods of manipulating cognitive load should still be investigated using similar measures of load.

One problem with the current study is that the various measures of cognitive load were somewhat intrusive and may have created an unnatural learning situation. Finding unobtrusive and valid measures of each type of cognitive load continues to be a challenge for multimedia researchers. However, this challenge is worth meeting because the concept of cognitive load plays a central role in most theories of instructional design.

### References

- Ayres, P. (2006). Using subjective measures to detect variations of intrinsic cognitive load within problems. *Learning and Instruction, 16*, 389–400.
- Brünken, R., Plass, J. L., & Luetner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist, 38*, 53–62.
- Brünken, R., Plass, J. L., & Leutner, D. (2004). Assessment of cognitive load in multimedia learning with dual-task methodology: Auditory load and modality effects. *Instructional Science, 32*, 115–132.
- Brünken, R., Steinbacher, S., Schnotz, W., Plass, J., & Leutner, D. (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. *Experimental Psychology, 49*, 109–119.
- Chandler, P., & Sweller, J. (1996). Cognitive load while learning to use a computer program. *Applied Cognitive Psychology, 10*, 151–170.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Kalyuga, S. (2005). Prior knowledge principle in multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 325–338). New York: Cambridge University Press.
- Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia learning. *Applied Cognitive Psychology, 13*, 351–371.
- Marcus, N., Cooper, M., & Sweller, J. (1996). Understanding instructions. *Journal of Educational Psychology, 88*, 49–63.
- Mayer, R. E. (2001). *Multimedia learning*. New York: Cambridge University Press.
- Mayer, R. E. (2005a). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 31–48). New York: Cambridge University Press.
- Mayer, R. E. (2005b). Principles for reducing extraneous processing in multimedia learning: Coherence, signaling, redundancy, spatial contiguity, and temporal contiguity principles. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 183–200). New York: Cambridge University Press.
- Mayer, R. E., & Chandler, P. (2001). When learning is just a click away: Does simple user interaction foster deeper understanding of multimedia messages? *Journal of Educational Psychology, 93*, 390–397.
- Mayer, R. E., Dow, G., Mayer, S. (2003). Multimedia learning in an interactive self-explaining environment: What works in the design of agent-based microworlds? *Journal of Educational Psychology, 95*, 806–812.
- Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*, 43–52.
- Paas, F., Tuovinen, J. E., Tabbers, H., & van Gerven, P. W. M. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*, 63–72.
- Paas, F., & van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review, 6*, 51–71.
- Paas, F., van Merriënboer, J. J., & Adam, J. J. (1994). Measurement of cognitive load in instructional research. *Perceptual & Motor Skills, 79*, 419–430.
- Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257–285.
- Sweller, J. (1999). *Instructional design in technical areas*. Camberwell, Victoria, Australia: ACER Press.
- Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 19–30). New York: Cambridge University Press.

Received December 12, 2006

Revision received August 22, 2007

Accepted September 4, 2007 ■

The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

**Manuscript preparation.** Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (5th ed.). Manuscripts may be copyedited for bias-free language (see chap. 2 of the *Publication Manual*). Formatting instructions (all copy must be double-spaced) and instructions on the preparation of tables, figures, references, metrics, and abstracts appear in the *Manual*. For APA's Checklist for Manuscript Submission, see [www.apa.org/journals](http://www.apa.org/journals). **Abstract and key-words.** All manuscripts must include an abstract containing a maximum of 180 words typed on a separate page. After the abstract, please supply up to five keywords or brief phrases. **References.** References should be listed in alphabetical order. Each listed reference should be cited in text, and each text citation should be listed in the References. Basic formats are as follows:

- Haag, L., & Stern, E. (2003). In search of the benefits of learning Latin. *Journal of Educational Psychology*, 95, 174–178.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Johnson, D. W., & Johnson, R. T. (1990). Cooperative learning and achievement. In S. Sharon (Ed.), *Cooperative learning: Theory and research* (pp. 173–202). New York: Praeger.

Adequate description of participants is critical to the science and practice of educational psychology; this allows readers to assess the results, determine generalizability of findings, and make comparisons in replications, extensions, literature reviews, or secondary data analyses. Authors should see guidelines for sample-subject description in the *Manual*. Appropriate indexes of effect size or strength of relationship should be incorporated in the results section of the manuscript (see pp. 5, 25–26 of the *Manual*). Information that allows the reader to assess not only the significance but also the magnitude of the observed effects or relationships clarifies the importance of the findings. **Figures.** Graphics files are welcome if supplied as Tiff, EPS, or PowerPoint. High-quality printouts or glossies are needed for *all* figures. The minimum line weight for line art is 0.5 point for optimal printing. When possible, please place symbol legends below the figure image instead of to the side. Original color figures can be printed in color at the editor's and publisher's discretion provided the author agrees to pay \$255 for one figure, \$425 for two figures, \$575 for three figures, \$675 for four figures, and \$55 for each additional figure.

**Publication policies.** APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. APA policy regarding posting articles on the Internet may be found at [www.apa.org/journals](http://www.apa.org/journals). In addition, it is a violation of APA Ethical Principles to publish "as original data, data that have been previously published" (Standard 8.13). As this is a primary journal that publishes original material only, APA policy prohibits publication of any manuscript or data that have already been published in whole or substantial part elsewhere. Authors have an obligation to consult journal editors concerning prior publication of any data on which their article depends. In addition, APA Ethical Principles specify that "after research

results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release" (Standard 8.14). Authors must have available their data throughout the editorial review process and for at least 5 years after the date of publication.

**Masked review policy.** Because the *Journal* has a masked review policy, authors submitting manuscripts are required to include, with each copy of the manuscript, a cover sheet that shows the title of the manuscript, the authors' names and institutional affiliations, the date the manuscript is submitted, and footnotes identifying the authors or their affiliations. The first page of the manuscript should omit the authors' names and affiliations but should include the title of the manuscript and the date it is submitted. Every effort should be made by the authors to see that the manuscript itself contains no clues to their identities. Authors are required to state in writing that they have complied with APA ethical standards in the treatment of their sample, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained at [www.apa.org/ethics/](http://www.apa.org/ethics/) or by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242. APA requires authors to reveal any possible conflict of interest in the conduct and reporting of research (e.g., financial interests in a test procedure, funding by pharmaceutical companies for drug research). Authors of accepted manuscripts will be required to transfer copyright to APA.

**Permissions.** Authors of accepted papers are required to obtain and provide to the editor on final acceptance all necessary permissions to reproduce in print and electronic form any copyrighted work, including, for example, test materials or portions thereof and photographs of people.

**Supplemental materials.** APA can now place supplementary materials online, which will be available via the published article in the PsycARTICLES database. To submit such materials, please see [www.apa.org/journals/authors/suppmaterial.html](http://www.apa.org/journals/authors/suppmaterial.html) for details. Authors of accepted papers will be asked to work with the editor and production staff to provide supplementary materials as appropriate.

**Submission.** Authors should submit their manuscripts electronically via the Manuscript Submission Portal at [www.apa.org/journals/edu](http://www.apa.org/journals/edu) (follow the link "Submit Manuscripts Electronically"). A checklist for manuscript submission, including guidelines for preparing the electronic file, can be found at [www.apa.org/journals/](http://www.apa.org/journals/). Correspondence regarding manuscripts should be sent to the Editor, Art Graessen, University of Memphis, Journal of Educational Psychology, 202 Psychology Building, Memphis, TN 38152-3230. In addition to addresses and phone numbers, authors should supply e-mail addresses, as most communications will be by e-mail. Fax numbers, if available, should also be provided for potential use by the editorial office and later by the production office. Authors should keep a copy of the manuscript to guard against loss. E-mail correspondence may be addressed to [jedgar@memphis.edu](mailto:jedgar@memphis.edu).

**Preparing files for production.** If your manuscript is accepted for publication, please follow the guidelines for file formats and naming provided at [www.apa.org/journals/authors/preparing\\_files.html](http://www.apa.org/journals/authors/preparing_files.html). If your manuscript was mask reviewed, please ensure that the final version for production includes a byline and full author note for typesetting.



# BEST SELLERS

from the American Psychological Association



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

## Unlocking the Potential of Patients With ADHD

A Model for Clinical Practice

*Vincent J. Monastra*

2008. 328 pages. Hardcover.

List: \$69.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0238-6 ■ Item # 4317139

## Medically Unexplained Illness

Gender and Biopsychosocial Implications

*Susan K. Johnson*

2008. 280 pages. Hardcover.

List: \$69.95

APA Member/Affiliate: \$49.95

ISBN 978-0-9792125-8-1 ■ Item # 4317135

## Pedophilia and Sexual Offending Against Children

Theory, Assessment, and Intervention

*Michael C. Seto*

2008. 304 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-4338-0114-3 ■ Item # 4317136

## Sex Offending

Causal Theories to Inform Research, Prevention, and Treatment

*Jill D. Stinson, Bruce D. Sales, and Judith V. Becker*

List: \$69.95

APA Member/Affiliate: \$49.95

ISBN 978-0-9792125-2-9 ■ Item # 4316094

## Directory of Unpublished Experimental Mental Measures

Volume 9

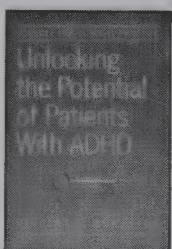
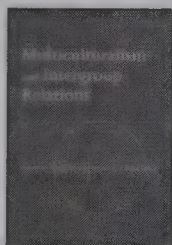
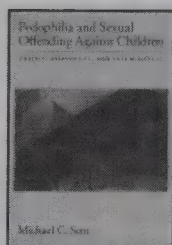
*Bert A. Goldman and David F. Mitchell*

2008. 832 pages. Hardcover.

List: \$69.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0137-2 ■ Item # 4316096



## Emotion-Focused Couples Therapy

The Dynamics of Emotion, Love, and Power

*Leslie S. Greenberg and Rhonda N. Goldman*

2008. 384 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0316-1 ■ Item # 4317147

## Behavioral Interventions in Cognitive Behavior Therapy

Practical Guidance

for Putting Theory Into Action

*Richard F. Farmer and Alexander L. Chapman*

2008. 280 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0241-6 ■ Item # 4317144

## Childhood Mental Health Disorders

Evidence Base and Contextual Factors for Psychosocial, Psychopharmacological, and Combined Interventions

*Ronald T. Brown, et. al.*

2008. 200 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0170-9 ■ Item # 4317137

## Polarities of Experience

Relatedness and Self-Definition in Personality Development, Psychopathology, and the Therapeutic Process

*Sidney J. Blatt*

2008. 424 pages. Hardcover.

List: \$69.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0314-7 ■ Item # 4317146

## Hallucinations

The Science of Idiosyncratic Perception

*André Aleman and Frank Larøi*

2008. 328 pages. Hardcover.

List: \$69.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0311-6 ■ Item # 4318044

## The Q-Sort in Character Appraisal

Encoding Subjective Impressions of Persons Quantitatively

*Jack Block*

2008. 208 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0315-4 ■ Item # 4316103

## Graduate Study in Psychology

2008 Edition

2008. 832 pages. Paperback.

List: \$25.95

APA Member/Affiliate: \$22.95

ISBN 978-1-4338-0128-0 ■ Item # 4270091

## Addressing Cultural Complexities in Practice

Second Edition

Assessment, Diagnosis, and Therapy

*Pamela A. Hays*

2008. 320 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0219-5 ■ Item # 4317138

## Multiculturalism and Intergroup Relations

Psychological Implications

for Democracy in Global Context

*Faithali M. Moghaddam*

2008. 280 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$44.95

ISBN 978-1-4338-0307-9 ■ Item # 4317145

## Inclusive Cultural Empathy

Making Relationships Central in Counseling and Psychotherapy

2008. 395 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-0-9792125-1-2 ■ Item # 4317133

## Listening to Battered Women

A Survivor-Centered Approach to Advocacy, Mental Health, and Justice

*Lisa A. Goodman and Deborah Epstein*

2008. 208 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$44.95

ISBN 978-1-4338-0239-3 ■ Item # 4317143

## Commemorating Brown

The Social Psychology of Racism and Discrimination

*Edited by Glenn Adams, Monica Biernat, Nyla R. Branscombe, Christian S. Crandall, and Lawrence S. Wrightsman*

2008. 416 pages. Hardcover.

List: \$69.95

APA Member/Affiliate: \$54.95

ISBN 978-1-4338-0308-6 ■ Item # 4316098

## Gender and Occupational Outcomes

Longitudinal Assessment of Individual, Social, and Cultural Influences

*Edited by Helen M. G. Watt*

*and Jacquelynne S. Eccles*

2008. 542 pages. Hardcover.

List: \$69.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0310-9 ■ Item # 4316102

## Self-Criticism and Self-Enhancement

Theory, Research, and Clinical Implications

*Edited by Edward C. Chang*

2008. 296 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0115-0 ■ Item # 4318043

AD0552

To Order: 800-374-2721 • [www.apa.org/books](http://www.apa.org/books)



# SHY CHILDREN, PHOBIC ADULTS

Nature and Treatment of Social Anxiety Disorder, Second Edition

Deborah C. Beidel and Samuel M. Turner

**T**his book describes the clinical presentation of social anxiety disorder, presents theoretical perspectives on its etiology, and examines the latest empirical data with respect to both pharmacological and behavioral interventions. Social anxiety disorder occurs in children, adolescents, and adults, but its manifestation and treatment differ depending on developmental factors. Drawing from a broad literature base as well as their extensive clinical experience, the authors illustrate the impact of developmental stage on all aspects of the disorder. They also provide practical implementation guidelines, enhanced by case examples, tips on patient management, lists of assessment instruments, and sample forms to use with clients.

Since publication of the first edition in 1998, knowledge about social anxiety disorder has advanced on several fronts. The new edition includes information from new studies differentiating patterns of distress characteristic of social anxiety disorder vs. social phobia. It draws on more substantive data bases to support firmer conclusions about the presentation of social anxiety disorder among children and adolescents as well as across various ethnocultural groups. New assessment strategies reviewed in this book include neuroassessments using magnetic resonance imaging, and well-validated self-report instruments and clinician rating scales. Authors review greatly expanded literature addressing pharmacological treatment and psychosocial treatments. New case descriptions and clinical materials are also included. This highly informative and comprehensive volume will be illuminating reading for practitioners, researchers, and students. 2007. 408 pages. Hardcover.

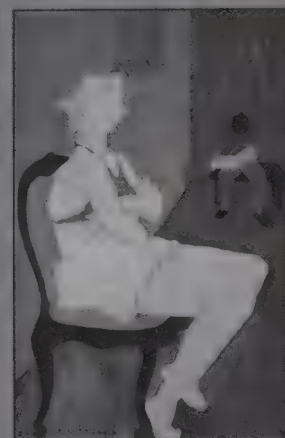
**CONTENTS:** Dedication ■ Acknowledgments ■ Preface ■ Introduction ■ Chapter 1. Clinical Presentation of Social Anxiety Disorder in Adults ■ Chapter 2. Clinical Presentation of Social Anxiety Disorder in Children and Adolescents ■ Chapter 3. Prevalence of Social Anxiety Disorder ■ Chapter 4. Etiology of Social Anxiety Disorder ■ Chapter 5. Assessment of Social Anxiety Disorder ■ Chapter 6. Managing Patients with Social Anxiety Disorder (and Their Parents) ■ Chapter 7. Pharmacological Treatment of Social Anxiety Disorder ■ Chapter 8. Behavioral and Cognitive-Behavioral Treatment of Social Anxiety Disorder in Adults ■ Chapter 9. Behavioral and Cognitive-Behavioral Treatment of Social Anxiety Disorder in Children and Adolescents ■ Epilogue ■ References ■ About the Authors

**List: \$59.95 • APA Member/Affiliate: \$49.95 • ISBN 1-59147-452-3 • Item # 4317118 • ISBN-13: 978-1-59147-452-4**

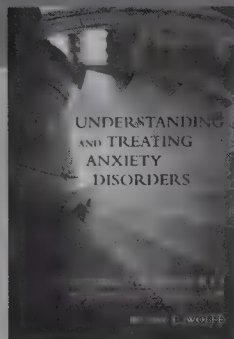
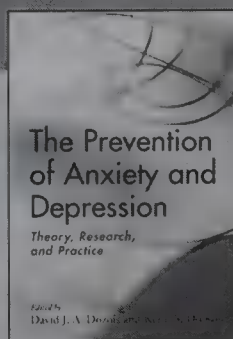
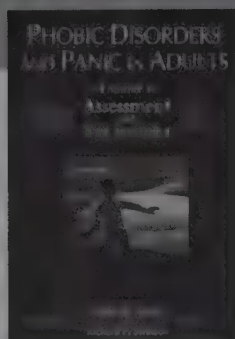
## Shy Children, Phobic Adults

SECOND EDITION

Nature and  
Treatment  
of Social  
Anxiety  
Disorder



Deborah C. Beidel  
Samuel M. Turner



ALSO AVAILABLE

**PHOBIC DISORDERS AND PANIC IN ADULTS: A Guide to Assessment and Treatment**  
Martin M. Antony and Richard P. Swinson

2000 • 422 pages ■ Hardcover • List: \$39.95 ■ APA Member/Affiliate: \$34.95  
ISBN 1-55798-696-7 ■ Item # 431756A ■ ISBN-13: 978-1-55798-696-2

**THE PREVENTION OF ANXIETY AND DEPRESSION: Theory, Research, and Practice**  
Edited by David J. A. Dozois and Keith S. Dobson

2004 ■ 330 pages ■ Hardcover ■ List: \$49.95 • APA Member/Affiliate: \$39.95  
ISBN 1-59147-079-X ■ Item # 4316024 • ISBN-13: 978-1-59147-079-3

**UNDERSTANDING AND TREATING ANXIETY DISORDERS**  
An Integrative Approach to Healing the Wounded Self

Barry E. Wolfe  
2005 ■ 301 pages ■ Hardcover • List: \$59.95 ■ APA Member/Affiliate: \$44.95  
ISBN 1-59147-196-6 ■ Item # 4317058 ■ ISBN-13: 978-1-59147-196-7

**APA Books**  
Ordering Information  
**800-374-2721**  
[www.apa.org/books](http://www.apa.org/books)

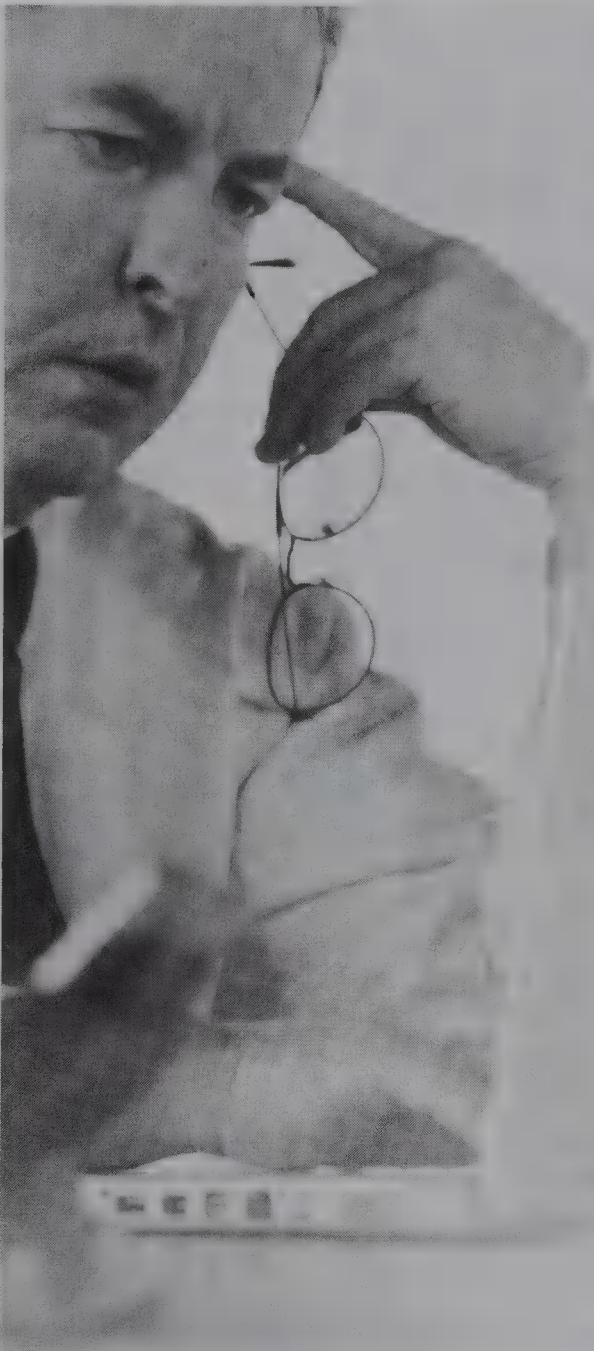
In Washington, DC, call: 202-336-5510  
TDD/TTY: 202-336-6123 • Fax: 202-336-5502  
In Europe, Africa, or the Middle East,  
call: 44-207-240-0856

AD0486



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION





## The Information You Need May Be on the Internet, But Can You Find It?

There's a vast universe of information available on the Internet today.

With such abundance, how do you get to the most critical and highest quality content you need for your research or studies? Many general search engines simply don't allow you to target reliable scholarly literature.

There's a simple solution —thanks to your institution's library.

When you are looking for specific resources from PubMed or APA journals, for example, log into your university library system first. You can search an enormous array of authoritative, professional information sources—such as APA's five electronic databases—because your library has purchased subscriptions. And while you're on the Web, you take your library's rights with you, giving you access to more sites and more content.

Plus, you can use your library's online gateway from your home or office to get the information you need ***free of charge*** and ***around the clock***.

Problem solved.



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

**[www.apa.org/publications](http://www.apa.org/publications)**

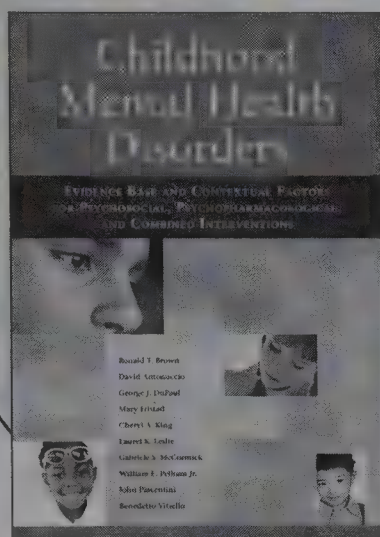
# Childhood Mental Health Disorders

## EVIDENCE BASE AND CONTEXTUAL FACTORS FOR PSYCHOSOCIAL, PSYCHOPHARMACOLOGICAL, AND COMBINED INTERVENTIONS

Ronald T. Brown, David O. Antonuccio, George J. DuPaul, Mary A. Fristad,  
Cberyl A. Kling, Laurel K. Leslie, Gabrielle S. McCormick,  
William E. Pelham, Jr., John C. Placentini,  
and Benedetto Vitiello



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION



**C**hildhood Mental Health Disorders is a comprehensive report, based on a thorough review of the literature, of the current recognized spectrum of childhood mental health disorders, including effective use of psychotropic medications and psychosocial interventions for children and adolescents. All of the studies come from top experts in the medical and psychology communities, and include the most recent, updated information, despite the rapidly changing face of child and adolescent mental health studies. Acknowledging the complexity of these disorders and the need to individualize treatment, the volume is intended as a basic yet comprehensive framework for anyone interested in mental health in children and adolescents. The disorders addressed include attention deficit/hyperactivity disorder, obsessive-compulsive disorder, autism and schizophrenia, and other commonly recognized disorders.

2008. 200 pages. Hardcover.  
ISBN 978-1-4338-0170-9 • Item # 4317137  
List: \$59.95 • APA Member/Affiliate: \$49.95

### Also Available



### Intervening in Children's Lives

*An Ecological, Family-Centered Approach to Mental Health Care*

Thomas J. Dishion and Elizabeth A. Stormshak

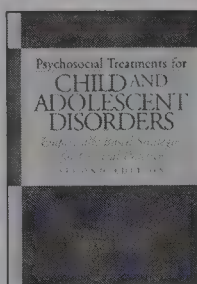
2007. 320 pages. Hardcover.  
ISBN 978-1-59147-428-9 • Item # 4317115  
List: \$69.95 • APA Member/Affiliate: \$49.95



### Chronic Health-Related Disorders in Children

*Collaborative Medical and Psychoeducational Interventions*  
Edited by LeAdelle Phelps

2006. 312 pages. Hardcover.  
ISBN 978-1-59147-408-1 • Item # 4317108  
List: \$59.95 • APA Member/Affiliate: \$49.95



### Psychosocial Treatments for Child and Adolescent Disorders

*Empirically Based Strategies for Clinical Practice*

Second Edition

Edited by Euthymia D. Hibbs and Peter S. Jensen

2005. 832 pages. Hardcover.  
ISBN 978-1-59147-092-2 • Item # 4318006  
List: \$69.95 • APA Member/Affiliate: \$54.95

### Contents

- Preface
- Foreword
- Chapter 1.** Introduction
- Chapter 2.** Attention-Deficit/Hyperactivity Disorder
- Chapter 3.** Oppositional-Defiant Disorder and Conduct Disorder
- Chapter 4.** Tourette and Tic Disorders
- Chapter 5.** Obsessive Compulsive Disorder
- Chapter 6.** Anxiety Disorders
- Chapter 7.** Depressive Disorders and Suicidality
- Chapter 8.** Bipolar Disorder
- Chapter 9.** Schizophrenia Spectrum Disorder
- Chapter 10.** Autism Spectrum Disorder
- Chapter 11.** Anorexia Nervosa and Bulimia Nervosa
- Chapter 12.** Elimination Disorders
- Chapter 13.** Future Directions References

### APA BOOKS ORDERING INFORMATION

**800-374-2721**

[www.apa.org/books](http://www.apa.org/books)

In Washington DC, call: 202-336-5510

TDD/TTY: 202-336-6123 Fax: 202-336-5502

In Europe, Africa, or the Middle East call: 44-1767-604972

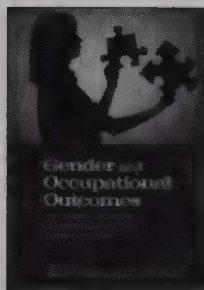


# NEW RELEASES

from the American Psychological Association



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION



## Gender and Occupational Outcomes

Longitudinal Assessment of Individual, Social, and Cultural Influences

*Edited by Helen M. G. Watt and Jacquelynne S. Eccles*

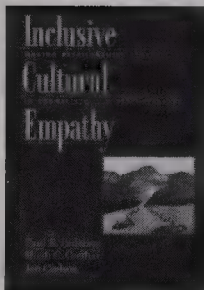
2008. 400 pages. Hardcover.

List: \$69.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0310-9

Item # 4316102



## Inclusive Cultural Empathy

Making Relationships Central in Counseling and Psychotherapy

*Paul B. Pedersen, Hugh C. Crethar, and Jon Carlson*

2008. 296 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-0-9792125-1-2

Item # 4317133



## Sex Offending

Causal Theories to Inform Research, Prevention, and Treatment

*Jill D. Stinson, Bruce D. Sales, and Judith V. Becker*

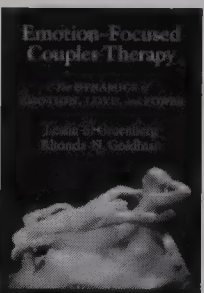
2008. 288 pages. Hardcover.

List: \$69.95

APA Member/Affiliate: \$49.95

ISBN 978-0-9792125-2-9

Item # 4316094



## Emotion-Focused Couples Therapy

The Dynamics of Emotion, Love, and Power

*Leslie S. Greenberg and Rhonda N. Goldman*

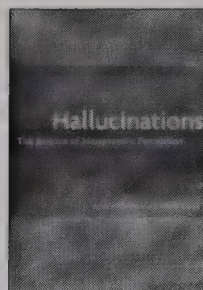
2008. 384 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0316-1

Item # 4317147



## Hallucinations

The Science of Idiosyncratic Perception

*André Aleman and Frank Larøi*

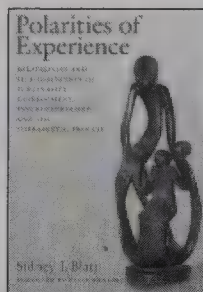
2008. 328 pages. Hardcover.

List: \$69.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0311-6

Item # 4318044



## Polarities of Experience

Relatedness and Self-Definition in Personality Development, Psychopathology, and the Therapeutic Process

*Sidney J. Blatt*

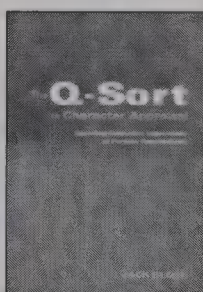
2008. 424 pages. Hardcover.

List: \$69.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0314-7

Item # 4317146



## The Q-Sort in Character Appraisal

Encoding Subjective Impressions of Persons Quantitatively

*Jack Block*

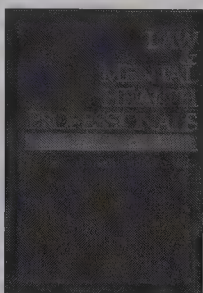
2008. 208 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0315-4

Item # 4316103



## Law and Mental Health Professionals

Kansas

*Jan Bowen Sheldon and Scott A. Letts*

2008. 528 pages. Hardcover.

List: \$99.95

APA Member/Affiliate: \$74.95

ISBN 978-1-4338-0331-4

Item # 4315009

AD0564

To Order: 800-374-2721 • [www.apa.org/books](http://www.apa.org/books)



Volume 100  
Number 2

May 2008

Published quarterly  
by the  
American Psychological  
Association

ISSN 0022-0663

# Journal of Educational Psychology

Barbara R. Harris, *Editor*

Eric M. Anderman, *Associate Editor*

Donna M. Kulikowich, *Associate Editor*

Moria Miller, *Associate Editor*

Frank Pajares, *Associate Editor*

Jeffrey J. Walczyk, *Associate Editor*

**CURRENT YR/VOL**

**Marygrove College Library  
8425 West McNichols Road  
Detroit, MI 48221**

[www.apa.org/journals/edu](http://www.apa.org/journals/edu)

2000-2010  
DECADE  
of BEHAVIOR



## Editor

Karen R. Harris, *Vanderbilt University*

## Associate Editors

Eric M. Anderman, *Ohio State University*  
Jonna M. Kulikowich, *Pennsylvania State University*  
Gloria Miller, *University of Denver*  
Frank Pajares, *Emory University*  
Jeffrey J. Walczyk, *Louisiana Technical University*

## Chief Editorial Assistant

Brenna Hansen, *Vanderbilt University*

## Editorial Assistants

Karrie Godwin, *University of Denver*  
Diana Griffith-Ross, *Louisiana Technical University*  
Jason Chen, *Emory University*  
Nicholas D. Warcholak, *Pennsylvania State University, University Park Campus*

## Advisory Editors

Patricia Alexander, *University of Maryland, College Park*  
Ellen R. Altermatt, *Hanover College*  
Lynley H. Anderman, *Ohio State University*  
Robert Atkinson, *Arizona State University*  
Carole Beal, *Information Sciences Institute at the University of Southern California*  
Hefer Bembenuddy, *Queens College*  
David A. Bergin, *University of Missouri—Columbia*  
Benita A. Blachman, *Syracuse University*  
Mimi Bong, *Ewha Womans University, Seoul, Korea*  
Jere Brophy, *Michigan State University*  
Scott W. Brown, *University of Connecticut*  
Adriana G. Bus, *Leiden University, Leiden, the Netherlands*  
Robert Calfee, *University of California, Riverside*  
Joanne F. Carlisle, *University of Michigan*  
Martha Carr, *University of Georgia*  
Jerrold C. Cassidy, *Ball State University*  
Clark Chinn, *Rutgers University*  
Namok Choi, *University of Louisville*  
Donald L. Compton, *Vanderbilt University*  
Alice J. Corkill, *University of Nevada, Las Vegas*  
H. Michael Crowson, *University of Oklahoma*  
Anne E. Cunningham, *University of California, Berkeley*  
Teresa K. DeBacker, *University of Oklahoma*  
Amanda M. Durik, *Northern Illinois University*  
Pamela Beard El-Dinary, *Educational Consultant*  
Dorothy L. Espelage, *University of Illinois at Urbana—Champaign*  
Jill Fitzgerald, *University of North Carolina at Chapel Hill*  
Douglas Fuchs, *Vanderbilt University*  
Lynn S. Fuchs, *Vanderbilt University*  
David C. Geary, *University of Missouri*  
Alexandra Gottardo, *Wilfrid Laurier University, Waterloo, Ontario, Canada*  
Steve Graham, *Vanderbilt University*  
Barbara A. Greene, *University of Oklahoma*  
Charles R. Greenwood, *University of Kansas*  
John Guthrie, *University of Maryland, College Park*  
Douglas J. Hacker, *University of Utah*  
Vernon C. Hall, *Syracuse University*  
Jenefer Husman, *Arizona State University*  
Michael L. Kamil, *Stanford University*  
Avi Kaplan, *Ben Gurion University of the Negev, Beer Sheva, Israel*  
Robert M. Klassen, *University of Alberta, Edmonton, Alberta, Canada*  
Beth Kurtz-Costes, *University of North Carolina at Chapel Hill*  
Dan Lapsley, *University of Notre Dame*  
Steve Lehman, *Utah State University*  
Willy Lens, *University of Leuven, Leuven, Belgium*  
Joel R. Levin, *University of Arizona*  
Elizabeth A. Linnenbrink, *Duke University*  
Mary Lundeberg, *Michigan State University*  
Charles MacArthur, *University of Delaware*  
Linda H. Mason, *Pennsylvania State University, University Park Campus*  
Richard E. Mayer, *University of California, Santa Barbara*  
Catherine McBride-Chang, *Chinese University of Hong Kong, Shatin, Hong Kong, China*  
Valentina Mcinerney, *University of Western Sydney*  
Debra K. Meyer, *Elmhurst College*  
Michael Middleton, *University of New Hampshire*  
Lisa M. Soederberg Miller, *University of California, Davis*  
Raymond B. Miller, *University of Oklahoma*  
Jens Möller, *University of Kiel, Kiel, Germany*  
Tamera B. Murdock, *University of Missouri—Kansas City*  
Karen P. Murphy, *Pennsylvania State University, University Park Campus*  
Darcia Narvaez, *University of Notre Dame*  
Markku Niemivirta, *University of Helsinki, Helsinki, Finland*  
Jane Oakhill, *University of Sussex, Falmer, Brighton, United Kingdom*  
Rollanda E. O'Connor, *University of California, Riverside*  
Richard Olson, *University of Colorado*  
Helen Patrick, *Purdue University*  
Nancy Perry, *University of British Columbia, Vancouver, British Columbia, Canada*  
Gary Phye, *Iowa State University*  
Jan L. Plass, *New York University*  
Robert Reid, *University of Nebraska—Lincoln*  
Robert Renaud, *University of Manitoba, Winnipeg, Manitoba, Canada*  
Alison M. Ryan, *University of Illinois at Urbana—Champaign*  
Hollis S. Scarborough, *Haskins Laboratories, New Haven, Connecticut*  
Christopher Schatschneider, *Florida State University*  
Wolfgang Schneider, *Universität Würzburg, Würzburg, Germany*  
Marlene Schommer-Aikins, *Wichita State University*  
Gregory Schraw, *University of Nevada, Las Vegas*

Einar M. Skaalvik, *Norwegian University of Science and Technology, Trondheim, Norway*  
Susan Sonnenschein, *University of Maryland, Baltimore County*  
Laura M. Stapleton, *University of Maryland, Baltimore County*  
Joseph Stevens, *University of Oregon*  
H. Lee Swanson, *University of California, Riverside*  
John Sweller, *University of New South Wales, Sydney, New South Wales, Australia*  
Sonya Symons, *Acadia University, Wolfville, Nova Scotia, Canada*  
Keith Thiede, *University of Illinois at Chicago*  
Theresa A. Thorkildsen, *University of Illinois at Chicago*  
Tim Urdan, *Santa Clara University*  
Ellen Usher, *University of Kentucky*  
Giovanni Valiante, *Rollins College*  
Sharon Vaughn, *University of Texas at Austin*  
Regina Vollmeyer, *University of Frankfurt, Frankfurt, Germany*  
Charles A. Weaver III, *Baylor University*  
Kathryn R. Wentzel, *University of Maryland, College Park*  
Allan Wigfield, *University of Maryland, College Park*  
Joanna P. Williams, *Teachers College, Columbia University*  
Christopher A. Wolters, *University of Houston*  
Moshe Zeidner, *University of Haifa, Haifa, Israel*  
Barry J. Zimmerman, *Graduate Center, City University of New York*

The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

**Change of Address:** Send change of address notice and a recent mailing label to the attention of Subscriptions Department, APA, 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee periodicals forwarding postage.

**Single Issues, Back Issues, and Back Volumes:** For information regarding single issues, back issues, or back volumes, write to Order Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

**Microform Editions:** For information regarding microform editions, write to University Microfilms, Ann Arbor, MI 48106.

**Manuscripts:** Submit manuscripts electronically through the Manuscript Submissions Portal found at [www.apa.org/journals/edu](http://www.apa.org/journals/edu), according to the Instructions to Authors found elsewhere in this issue (see table of contents). Correspondence regarding manuscripts should be sent to the Editor, Karen R. Harris, Vanderbilt University, *Journal of Educational Psychology*, Box 507 Peabody College, Nashville, Tennessee 37203-5721. The opinions and statements published are the responsibility of the authors, and such opinions and statements do not necessarily represent the policies of APA or the views of the Editor.

**Copyright and Permission:** Those who wish to reuse APA-copyrighted material in a non-APA publication must secure from APA and the author of the material written permission to reproduce a journal article in full or journal text of more than 500 words. APA normally grants permission contingent on permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of a fee of \$25 per page. Permission from APA and fees are waived for those who wish to reproduce a single table or figure from a journal for use in a print product, provided the author's permission is obtained and full credit is given to APA as copyright holder and to the author through a complete citation. (Requesters requiring written permission for commercial use of a single table or figure will be assessed a \$25 service fee.) Permission and fees are waived for authors who wish to reproduce their own material for personal use; fees only are waived for authors who wish to use more than a single table or figure of their own material commercially (but for use in edited books, fees are waived for the author only if serving as the book editor). Permission and fees are waived for the photocopying of isolated journal articles for nonprofit classroom or library reserve use by instructors and educational institutions. A permission fee may be charged to the requester if students are charged for the material, multiple articles are copied, or large-scale copying is involved (e.g., for course packs). Access services may use unedited abstracts without the permission of APA or the author. Libraries are permitted to photocopy beyond the limits of the U.S. copyright law: (1) post-1977 articles, provided the per-copy fee in the code for this journal (0022-0663/08/\$12.00) is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923; (2) pre-1978 articles, provided that the per-copy fee stated in the Publishers' Fee List is paid through the Copyright Clearance Center. Address requests for reprint permission to Permissions Office, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

**Electronic Access:** APA members who subscribe to this journal have automatic access to a 3-year file of the journal in the PsycARTICLES® full-text database. See <http://member-sapa.org/access>.

**Reprints:** Authors may order reprints of their articles from the printer when they receive proofs.

**APA Journal Staff:** Susan J. A. Harris, *Senior Director, Journals Program*; Skip Maier, *Director, Journal Services*; Paige W. Jackson, *Director, Editorial Services*; Clark Munsell, *Account Manager*; Jodi Ashcraft, *Director, Advertising Sales and Exhibits*.

The **Journal of Educational Psychology** (ISSN 0022-0663) is published quarterly (February, May, August, November) in one volume per year by the American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Subscriptions are available on a calendar year basis only (January through December). The 2008 rates follow: *Nonmember Individual*: \$161 Domestic, \$185 Foreign, \$195 Air Mail. *Institutional*: \$450 Domestic, \$491 Foreign, \$504 Air Mail. *APA Member*: \$73. Write to Subscriptions Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Printed in the U.S.A. Periodicals postage paid at Washington, DC, and at additional mailing offices. POSTMASTER: Send address changes to the *Journal of Educational Psychology*, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

# Educational Psychology

[www.apa.org/journals/edu](http://www.apa.org/journals/edu)

May 2008

Volume 100  
Number 2

## Articles

Copyright © 2008  
by the  
American  
Psychological  
Association

- 235 Socioeconomic Differences in Reading Trajectories: The Contribution of Family, Neighborhood, and School Contexts  
*Nikki L. Aikens and Oscar Barbarin*
- 252 Preschool Home Literacy Practices and Children's Literacy Development: A Longitudinal Analysis  
*Michelle Hood, Elizabeth Conlon, and Glenda Andrews*
- 272 Repeated Reading Intervention: Outcomes and Interactions With Readers' Skills and Classroom Instruction  
*Patricia F. Vadasy and Elizabeth A. Sanders*
- 291 On Warm Conceptual Change: The Interplay of Text, Epistemological Beliefs, and Topic Interest  
*Lucia Mason, Monica Gava, and Angela Boldrin*
- 310 Relationships of Three Components of Reading Fluency to Reading Comprehension  
*Susan Lutz Klauda and John T. Guthrie*
- 322 Native Language Proficiency, English Literacy, Academic Achievement, and Occupational Attainment in Limited-English-Proficient Students: A Latent Growth Modeling Perspective  
*R. Sergio Guglielmi*
- 343 Growth in Working Memory and Mathematical Problem Solving in Children at Risk and Not at Risk for Serious Math Difficulties  
*H. Lee Swanson, Olga Jerman, and Xinhua Zheng*
- 380 Revising the Redundancy Principle in Multimedia Learning  
*Richard E. Mayer and Cheryl I. Johnson*
- 387 Does Extrinsic Goal Framing Enhance Extrinsic Goal-Oriented Individuals' Learning and Performance? An Experimental Test of the Match Perspective Versus Self-Determination Theory  
*Maarten Vansteenkiste, Tinneke Timmermans, Willy Lens, Bart Soenens, and Anja Van den Broeck*
- 398 Task Values, Achievement Goals, and Interest: An Integrative Analysis  
*Chris S. Hulleman, Amanda M. Durik, Shaun A. Schweigert, and Judith M. Harackiewicz*
- 417 Striving for Social Dominance Over Peers: The Implications for Academic Adjustment During Early Adolescence  
*Sarah M. Kiefer and Allison M. Ryan*
- 429 Do Peers Contribute to the Likelihood of Secondary School Graduation Among Disadvantaged Boys?  
*Marie-Hélène Véronneau, Frank Vitaro, Sara Pedersen, and Richard E. Tremblay*

(Contents continue)



- 443 Control, Motivation, Affect, and Strategic Self-Regulation in the College Classroom: A Multidimensional Phenomenon  
*Duane F. Shell and Jenefer Husman*
- 460 What Makes Lessons Interesting? The Role of Situational and Individual Factors in Three School Subjects  
*Yi-Miau Tsai, Mareike Kunter, Oliver Lüdtke, Ulrich Trautwein, and Richard M. Ryan*
- 473 Cognitive Processing About Classroom-Relevant Contexts: Teachers' Attention to and Utilization of Girls' Body Size, Ethnicity, Attractiveness, and Facial Affect  
*Shirley S. Wang, Teresa A. Treat, and Kelly D. Brownell*
- 

## Other

- 416 American Psychological Association Subscription Claims Information
- 442 E-Mail Notification of Your Latest Issue Online!
- 490 Instructions to Authors
- 386 Low Publication Prices for APA Members and Affiliates
- 321 Members of Underrepresented Groups: Reviewers for Journal Manuscripts Wanted
- 309 Subscription Order Form

# Socioeconomic Differences in Reading Trajectories: The Contribution of Family, Neighborhood, and School Contexts

Nikki L. Aikens  
Mathematica Policy Research, Inc.

Oscar Barbarin  
University of North Carolina at Chapel Hill

In the present study, the authors use the Early Childhood Longitudinal Study, Kindergarten Cohort of 1998–1999, to examine the extent to which family, school, and neighborhood factors account for the impact of socioeconomic status (SES) on children's early reading. Through the use of hierarchical linear modeling techniques, growth curve models were estimated to depict children's reading trajectories from kindergarten to 3rd grade. Family characteristics made the largest contribution to the prediction of initial kindergarten reading disparities. This included home literacy environment, parental involvement in school, and parental role strain. However, school and neighborhood conditions contributed more than family characteristics to SES differences in learning rates in reading. The association between school characteristics and reading outcomes suggests that makeup of the student population, as indexed by poverty concentration and number of children with reading deficits in the school, is related to reading outcomes. The findings imply that multiple contexts combine and are associated with young children's reading achievement and growth and help account for the robust relation of SES to reading outcomes.

*Keywords:* socioeconomic status, reading, achievement trajectories

Socioeconomic status (SES) differences in children's reading and educational outcomes are ubiquitous, stubbornly persistent, and well documented (Arnold & Doctoroff, 2003; Duncan, Yeung, Brooks-Gunn, & Smith, 1998; Lee & Burkam, 2002; McLoyd, 1998; Yeung, Linver, & Brooks-Gunn, 2002). Economically disadvantaged children acquire language skills more slowly, exhibit delayed letter recognition and phonological sensitivity, and are at risk for reading difficulties (Whitehurst & Lonigan, 1998). Speculation about the mechanisms through which economic status affects reading achievement often concerns family life, and more recently, contexts like schools and neighborhoods are promising sources of influence. In the present study, we examine the contribution of these multiple settings to SES differences in early reading. This study is guided by ecological and developmental systems theories, in which it is recognized that children's lives unfold within multiple settings and in relationships with multiple others. Each of these settings functions in a dynamic, reciprocal process between the settings and the individual child (Bronfenbrenner, 1979, 1989; Magnusson & Cairns, 1996). Under such perspectives, functioning and development are not merely reflections of children themselves but also of the nature of experiences, resources, and

interactions encountered by children across settings. Functioning and development are also a reflection of how settings interact (see, for example, Early Child Care Research Network, National Institute of Child Health and Human Development [NICHD], 2004). Specifically, these frameworks account for influences of social environment on reading development in terms of (a) the qualities of environments, such as climate, activities, resources, and strains, and (b) the quality of social relations within and across these settings. Through their resources, experiences, and interactions, families, schools, and neighborhoods create auspicious or risky environments for children's reading development.

Social relationships include parent–child, teacher–child, and peer–peer interactions. The nature and quality of interactions with important adults are important to children's academic and social–emotional development (NICHD, 1998, 2000). Maternal warmth and mother–child interaction patterns have been cited as important to children's language outcomes (Hess, Holloway, Dickson, & Price, 1984; Murray & Hornbaker, 1997).

Bronfenbrenner's (1979, 1989) work recognized that systems differ in their proximity to the child, with more distal settings emanating outward from the child. The framework also captures the interrelations of systems. For example, parents' reports of involvement with the school represent the interaction of two settings—the family and the school. Here, we focus on those settings that children have direct contact with, including family, neighborhood, and school characteristics. We recognize that each of these settings may produce unique and cumulative effects on early development. Questions remain about the relative contribution of these contexts to the development of reading and to the relation of SES to early literacy outcomes. Ecological and developmental systems frameworks also suggest ways environments may interact with one another and may thus attenuate or amplify their effects on development. How, for instance, do family, neighborhood, and school contexts combine to channel the effects of SES on reading

---

Nikki L. Aikens, Mathematica Policy Research, Inc.; Oscar Barbarin, School of Social Work, University of North Carolina at Chapel Hill.

This research was supported by a grant from the American Educational Research Association (AERA), which receives funds for its AERA Grants Program from the National Science Foundation and the National Center for Education Statistics of the Institute of Education Sciences (U.S. Department of Education) under National Science Foundation Grant REC-9980573. Opinions stated are those of Nikki L. Aikens and Oscar Barbarin and are not necessarily those of the granting agencies.

Correspondence concerning this article should be addressed to Nikki L. Aikens, Mathematica Policy Research, Inc., 600 Maryland Avenue, S. W., Suite 550, Washington, DC 20024. E-mail: [naikens@mathematica-mpr.com](mailto:naikens@mathematica-mpr.com)



development? Because of the relation between early reading outcomes and later academic success, research exploring the contribution of multiple settings to reading outcomes is critical in the effort to deconstruct and identify the processes that give rise to socioeconomic differences in achievement. The theory also recognizes that the relative salience of contexts may shift over time.

Although ecological and developmental system frameworks underscore the need to examine multiple contexts on developmental outcomes (Bronfenbrenner, 1979, 1989; Magnusson & Cairns, 1996), family environment has most often been explored because it is considered to be the principal contributor to differences in early literacy and language development associated with SES. Developmental research is replete with studies examining the relation between child outcomes and family climates of low-SES children. For example, this body of work has demonstrated that children in low-SES households have less exposure to books at home (Evans, 2004; Lee & Burkam, 2002; Vernon-Feagans, Hammer, Miccio, & Manlove, 2002; Whitehurst & Lonigan, 1998) and have parents who are less involved in their schooling (Evans, 2004). These children are less likely to be regularly read to by parents (Federal Interagency Forum on Child and Family Statistics, 2005; Lee & Burkam, 2002; Whitehurst & Lonigan, 1998). Each of these family climate indicators has been cited as a factor underlying disparities in early literacy and language outcomes. With respect to relationship indicators, maternal sensitivity also has been implicated in the association between SES and children's early language abilities (Raviv, Kessenich, & Morrison, 2004). In addition, evidence suggests that differences in the quality of parents' behaviors during joint book reading (Whitehurst & Lonigan, 1998) and in the frequency and the quality of language interaction with parents in the home (Hart & Risley, 1995; National Research Council, 2000) contribute to disparities in early reading-related outcomes.

Only recently has there been an investigation of the influence of more distal settings, including those outside the home, on such reading-related differences. More important, few studies have concurrently examined the effects of multiple contexts on children's learning outcomes (see NICHD, 2004; Weigel, Martin, & Bennett, 2005, for recent exceptions). Exploring the contribution of these additional settings is important because interpreting SES effects as emanating exclusively from the family or the child means that policy and program interventions may focus too narrowly as they attempt to improve the educational outcomes of low-SES children. In addition, by neglecting the interactions of multiple contexts, one risks missing the dynamic and contextual nature of development and its outcomes (Cairns, Elder, & Costello, 1996).

Important differences in classroom, teacher, school, and community resources emerge along socioeconomic lines (Evans, 2004; Kozol, 1991; Lee & Burkam, 2002; Rothstein, 2004), highlighting the fact that socioeconomic disadvantage exposes poor children to multiple risky extrafamilial environments, including low quality child care, poor and distressed schools, and economically depressed neighborhoods. Given the confluence of negative factors associated with socioeconomic disadvantage, research examining the influence of multiple, nested settings may capture more accurately and more fully the influences that shape children's reading and academic growth than do single factor explanations.

Factors that put children at risk for learning and school difficulty, particularly those related to socioeconomic background, are distributed across environments and are not restricted to a single setting (Evans, 2004; Goldenberg, 2002; Leventhal & Brooks-

Gunn, 2000; Pianta & Walsh, 1996). For example, there is a strong association among school quality, classroom quality, and family SES (Entwisle & Alexander, 1993; Pianta, LaParo, Payne, Cox, & Bradley, 2002), particularly in regard to teacher and peer quality, instructional methods, and physical and material resources (Evans, 2004; Kozol, 1991; Lee & Burkam, 2002; Rothstein, 2004). Child and school poverty have also been linked to kindergarten teachers' positive interactions, classroom's instructional climate, and classroom's child-centered climate (Pianta et al., 2002). Differences in classroom quality as early as preschool are connected to early reading-related outcomes (Bryant, Burchinal, Lau, & Sparling, 1994; Dickinson, 2001; NICHD Early Child Care Research Network, 2000). A plethora of studies have assessed the relation of preschool programs, elementary school reading instruction, and peers to reading competence (see NICHD, 2004; Snow, Burns, & Griffin, 1998; Whitehurst & Lonigan, 1998; Xue & Meisels, 2004). This work suggests that classroom environments that are rich in literacy materials, that have teachers with high expectations of students and with adequate preparation to teach reading, that provide opportunities for dialogic reading or for children to be involved in the book reading experience, that provide support and opportunities for writing, and that promote stimulating teacher-child conversations enhance early language and literacy skills. This work also suggests that peers play a role in influencing early reading proficiency, with children in schools with larger concentrations of less skilled, lower SES, and minority peers exhibiting lower gains in reading during the kindergarten year (Xue & Meisels, 2004). However, remarkably few studies have sought to identify the contribution of school characteristics to the relations between SES and children's reading achievement.

Compared with more advantaged peers, economically disadvantaged children are more likely to reside in poorer quality neighborhoods (Evans, 2004; Lee & Burkam, 2002). Although research has highlighted mechanisms by which neighborhoods and communities impact adolescent development, especially behavioral problems, very few of these studies have explored reading outcomes, and research on very young children is rare. What has been done links neighborhood quality, children's verbal ability, IQ scores, and school achievement among young and early school age children (Leventhal & Brooks-Gunn, 2000). Again, however, in this work children's early reading outcomes have not been explicitly examined. These researchers have also not sought to understand the neighborhood characteristics that mediate the relations between socioeconomic background and children's reading achievement.

The current research assessed, among a nationally representative sample of children, the relative contribution of families, schools, and neighborhoods to the relation between SES and children's early reading achievement and growth between kindergarten and third grade. In addition, we sought to identify the predictors that account for this association across contexts and to identify how these relations vary over the first 4 years of school. This research assessed the extent to which the effects of SES on young children's reading development occur through what happens in the school, the neighborhood, and the home. Variables were selected to reflect those that have been previously linked in the literature to reading or academic achievement, with an emphasis on those that may differ across class lines. Accordingly, the present study is designed to answer the following questions:



1. What is the relation between SES and children's reading development from the fall of kindergarten to the spring of the third grade school year?
2. To what extent do resources, experiences, and interactions within family, neighborhood, and school contexts independently and cumulatively explain socioeconomic differences in early reading outcomes and growth? How do these relations vary over the first 4 years of school?

### Method

The data for the present study are from the Early Childhood Longitudinal Study, Kindergarten Class of 1998–1999 (ECLS-K; U.S. Department of Education, National Center for Education Statistics, 2004). The ECLS-K study followed a nationally representative sample of American children from kindergarten through fifth grade, providing a comprehensive picture of children's early family, neighborhood, and school experiences over time. The total sample consists of 21,260 children recruited from 1,277 kindergarten classrooms in public, Catholic, and non-Catholic private schools nationwide. A subsample of 17,401 children was retained in the longitudinal kindergarten to third grade sample. The study design involved multistage probability sampling to achieve a nationally representative sample of children eligible to attend kindergarten in the 1998–1999 school year.

### Instrumentation

During kindergarten, first-, and third-grade school years, the field staff conducted direct child assessments, parental interviews, and observations of the school environment and the community surrounding the school. Teachers and school administrators completed questionnaires during the spring data collection periods in kindergarten, first grade, and third grade. Data for the current analyses come from child, parent, teacher, school administrator, and field staff responses gathered during this time. A description of the items used in the current project follows.

*Child individual reading assessment.* Reading assessments involved a multiple choice or an open-ended format to measure basic skills (including print familiarity, letter recognition, mastery of beginning and ending sounds and rhyming sounds, word recognition), vocabulary (receptive vocabulary), and comprehension (listening comprehension, words in context). Test specifications were based on the 1992 and 1994 National Assessment of Educational Progress framework. The National Assessment of Educational Progress framework was developed to begin at fourth grade; consequently, the framework focuses on skills that are just emerging in early readers. For this reason, early elementary school educators and literacy curriculum specialists helped modify the assessment to be suitable for kindergarten and first grade. Pools of items for the reading assessment were borrowed or adapted, with permission, from published tests, including the Peabody Individual Achievement Test–Revised (Markwardt, 1989), the Peabody Picture Vocabulary Test–Revised (Dunn & Dunn, 1981), the Primary Test of Cognitive Skills (Huttenlocher & Levine, 1990), the Test of Early Reading Ability (Reid, Hresko, & Hammill, 1981), and the Woodcock-Johnson Tests of Achievement–Revised (Woodcock & Bonner, 1989).

Prior to administering the cognitive assessment, the Oral Language Development Scale (Duncan & DeAvila, 1986) was administered to those children identified from their school records (or by their teacher, if no school records were available) as coming from a household whose primary language was not English. This screening test determined whether a child was able to understand and respond to the cognitive assessment items in English. Children who passed the language screener received the full ECLS-K direct assessment battery. Children who did not pass an established cut score on the language screener did not receive the standard reading assessment.

The direct reading assessment consisted of a set of two-stage assessments: a first-stage routing section, followed by several alternative second-stage forms. First, a screening test consisting of 20 items with a broad range of difficulty was used to assign each child to one of three extensive batteries in the second stage of the assessment, in which items varied in difficulty. This was done to reduce testing time and to ensure that children would be required to respond to questions of moderate difficulty. Testing on this second-stage battery continued until the child reached a ceiling of three consecutive items missed. Item response theory (IRT) was used to generate standard scores based on an analysis of the item difficulty and the pattern of right and wrong responses. IRT scale scores allow for longitudinal measurement of achievement gain over time because assessments at each time are measured in a common metric and are therefore comparable across waves. IRT scores from each wave were used in the study's analyses as repeated measures of children's reading outcomes.

*Child variables.* The ECLS-K provides information on child gender, race, age at first assessment, and SES. Field staff identified child gender during the direct child assessment, and this information was subsequently confirmed during the parent interview (1 = female, 0 = male). A child composite for race provided eight categories for race and/or ethnicity. Children of all races are included in the current analyses, but a dichotomous, dummy variable for White children is modeled as a predictor of reading development. Child age reflects the child's age in months at the time of the fall kindergarten assessment. SES is a composite of five pieces of information: father's or male guardian's education, mother's or female guardian's education, father's or male guardian's occupation, mother's or female guardian's occupation, and household income. The index conceptualizes SES in quintiles, with lower scores indicating lower SES. SES is used in the current analyses because it tends to be less volatile than family income and encompasses several dimensions of economic conditions and parental capital beyond income.

*Family variables.* Parent interviews were completed primarily via telephone, and provisions were made to interview parents speaking Spanish, Lakota, Hmong, and Chinese. Parents provided information on child characteristics, family practices and resources, parent well-being, and home learning environment. Parents also reported the number of books the child had in the home (including library books). Data on whether the child had attended a day care center, nursery school, preschool, or prekindergarten program on a regular basis the year before he or she started kindergarten (1 = yes, 0 = no) were used as a measure of whether the child had received center-based care.

Three variables for the current analyses were derived from parents' responses. Home literacy environment is a sum score of



the frequency with which parents engaged in joint book reading with the child, the frequency with which children read (or pretended to read) books outside of school, and the frequency with which household members visited the library with the child. Parents' reports of the frequency of joint reading with the child and the frequency with which the child read (or pretended to read) had the same 4-point scale: 1 = *not at all*, 2 = *once or twice a week*, 3 = *three to six times a week*, and 4 = *everyday*. Library visits with the child within the last month was a dichotomous variable: 1 = yes, 0 = no.

Involvement in child's school is based on the sum of parents' reports of their own or another adult household member's participation in six activities reflecting involvement in the child's school (1 = yes, 0 = no), including attending a parent-teacher conference, attending a parent-teacher association (PTA) meeting, attending an open house, volunteering, participating in fundraising, and attending a school event (kindergarten  $\alpha = .58$ ; first-grade  $\alpha = .61$ ; third-grade  $\alpha = .61$ ). Parental role strain is represented by the average of parents' responses to nine items that assessed degree of difficulty and strain that the respondent experienced in his or her functioning as a parent (1 = *not at all true*, 2 = *somewhat true*, 3 = *mostly true*, 4 = *completely true*). Only four items were asked in the first grade, so an average of kindergarten and third grade data was estimated for first grade (kindergarten  $\alpha = .69$ ; third-grade  $\alpha = .56$ ). Conversely, parental warmth is represented by the average of parents' responses to four items assessing the warmth and closeness experienced in their relationship with the child (1 = *not at all true*, 2 = *somewhat true*, 3 = *mostly true*, 4 = *completely true*). Again, an average of kindergarten and third grade data was estimated for first grade (kindergarten  $\alpha = .54$ ; third-grade  $\alpha = .68$ ).

**Neighborhood variables.** Appraisals of the neighborhood setting were provided by three independent sources: parents, school administrators, and field staff. Parent reports of home neighborhood safety were assessed with a single item in which parents were asked about how safe it was to play in the neighborhood (1 = *not at all safe*, 2 = *somewhat safe*, 3 = *very safe*). Home neighborhood problems is a derived variable created from the sum of parents' responses about five qualities of their neighborhood, including the degree to which the following activities were a problem: garbage or litter in the street, individuals selling or using drugs in the street, burglary or robbery in the area, violent crime in the area, or vacant homes in the area (0 = *no problem*, 1 = *somewhat of a problem*, 2 = *big problem*). The average of the kindergarten and third-grade data was estimated for the first grade (kindergarten  $\alpha = .76$ ; third-grade  $\alpha = .75$ ).

School administrator reports on whether the community the school served was supportive of its goals and activities (1 = *strongly disagree*, 2 = *disagree*, 3 = *neither agree nor disagree*, 4 = *agree*, 5 = *strongly agree*) served as an index of community support for learning. A sum score of the degree to which school administrators viewed eight issues in the community surrounding the school as problems (tension from differences, unkempt areas, substance abuse, gangs, heavy traffic, violent crime, vacant buildings, and crime in the area; 0 = *no problem*, 1 = *somewhat of a problem*, 2 = *big problem*) represents school neighborhood problems (kindergarten  $\alpha = .82$ ; first-grade  $\alpha = .68$ ; third-grade  $\alpha = .76$ ).

Field staff observed and reported the presence of four conditions in the community near the school: litter and trash, graffiti,

boarded-up buildings, and people congregating (1 = *none*, 2 = *a little*, 3 = *some*, 4 = *a lot*). The average across the four items was used as the index of bad conditions near the school. The average of kindergarten and first-grade data was estimated for third grade (kindergarten  $\alpha = .95$ ; first-grade  $\alpha = .93$ ). Finally, the ECLS-K defines school urbanicity on the basis of census tract region: central city; large town, urban fringe; and small town, rural. The current analyses define schools located in the central city as urban; schools in large towns and urban fringe as suburban; and schools in small towns and rural areas as rural. A dichotomous, dummy variable for suburban is modeled as a predictor of reading development.

**School variables.** Classroom teachers completed self-administered questionnaires, providing reports on children's classroom peers, classroom literacy instruction, and teacher background and beliefs. School administrators also completed questionnaires reporting on the percentage of students in the school who were eligible to receive free or reduced price lunch. School poverty status was determined by supplementing this information with school-wide Title I status (at least 50% of the student body is poor), as reported by school administrators. In the case of missing data on the free or reduced price lunch variable, schools implementing school-wide Title I programs were identified as high-poverty schools. Consequently, a dichotomous variable reflecting high- (above 50%) and low-poverty (below 50%) schools was generated. School administrators also indicated whether the child's school was public or private. From this, a dichotomous, dummy variable for private school was generated for the current analyses.

Peers reading below grade is based on teachers' reports of the number of children reading below grade level in the classroom. Teacher experience is the sum of the number of years the teacher had taught at that school and the number of years the teacher had taught at that grade level. Teacher preparation is a variable derived from the sum of five teacher responses concerning their preparation to teach: the number of courses taken on early education, elementary education, and child development; the highest degree earned; and the number of courses taken on teaching reading (kindergarten  $\alpha = .62$ ; first-grade  $\alpha = .61$ ; third-grade  $\alpha = .60$ ).

Teachers reported the frequency of children's participation in several literacy-related activities in the classroom weekly, including working on learning the names of letters, practicing reading aloud and silently, reading a variety of texts, engaging in writing activities, and working on phonics. The number of activities reported varied at each grade level from 10 activities in kindergarten to 14 activities and 6 activities in first and third grades, respectively. Given differences in the scale across waves, items were rescaled to create a common scale: 0 = *never or hardly ever*, 1 = *monthly*, 2 = *weekly*, 3 = *daily or almost daily*. A sum score of these items was calculated across waves and represents literacy instruction (kindergarten  $\alpha = .61$ ; first-grade  $\alpha = .73$ ; third-grade  $\alpha = .69$ ).

### Analytic Sample

Exactly 10,998 of the children sampled in the fall kindergarten frame and followed to the spring third grade frame had positive, nonzero values on the appropriate sampling weight and were retained in the analytic sample. Thus, the present analyses are based on a weighted sample size of 3,842,954 children. Table 1

Table 1  
Percentage Distributions for Selected Characteristics of the Analytic Sample

Population	Total
Child race or ethnicity	
White	58%
Black	16%
Hispanic	19%
Asian	4%
Other	5%
Child sex	
Male	52%
Female	48%
Primary home language	
English	89%
Non-English	12%
Household income	
Below poverty line	21%
At or above poverty line	79%

Note.  $N = 3,842,954$ . Detail sums may not equal totals because of rounding.

presents the demographic characteristics of those whose data are represented in the current analyses.

### Analysis Overview

In the present study, AM (American Institutes of Research & Cohen, 2005; <http://am.air.org/default.asp>), a data analysis package designed to analyze complex sample survey data, was used for estimation of frequency and descriptive information. Because the ECLS-K data are not simple random samples, the analyses took into account the stratification of the survey design, and the study's sampling weights were used when computing frequency and descriptive information as well as tests of statistical significance. Standard errors were adjusted with a Taylor series approach. Use of sampling weights accounts for unequal probabilities of selection and nonresponse, and it also allows inferences from the analyses to be extrapolated to the larger population which the data were meant to represent. For the present analyses, longitudinal child sample weights from the ECLS-K kindergarten to third grade data file were used (C15PW0; this included children with parent interview data across the waves of interest in the analyses).

In addition, because the present study concerned the influence of family, neighborhood, and school variables on children's reading trajectories, growth curve models were estimated with hierarchical linear modeling (HLM) techniques. Growth curve modeling allows for the estimation of achievement trajectories that model change over time in relation to multiple influences. These models support investigation of covariates or predictors that underlie initial reading achievement, growth over time, and achievement outcomes at specific points in time during the period of interest. In the present study, four repeated measures of reading skills are modeled as a developmental trajectory. Specific child demographic characteristics (e.g., gender, SES, race, age) influence children's reading status in the fall of kindergarten and rates of reading growth through third grade. Additional family, neighborhood, and school characteristics help account for variation in children's initial reading, growth through third grade, and performance at specific points

in time beyond what can be accounted for by children's underlying trajectories. The HLM 6 software (Raudenbush, Bryk, & Congdon, 2005) was used to fit the multilevel model. The HLM program allows for the use of sampling weights that are associated with a complex sampling design and permits the estimation of models with random intercepts and slopes. It also permits nonequidistant times of observation, which is relevant in the ECLS-K data.

The data of all models were weighted at the child level in an effort to account for unequal probabilities of selection and nonresponse. The present HLM model has a general three-level nested structure: time at Level 1; child, family, and neighborhood characteristics at Level 2; and school characteristics at Level 3. Initial examination of the smoothed reading trajectories suggested that reading growth was not linear. As exhibited in Figure 1, many children exhibited a spike or growth spurt in reading scores between the spring of kindergarten and the spring of first grade, with slower reading growth exhibited before and after this period. Accordingly, a three-level model with intercept, three linear slope terms (SLOPE K: fall of kindergarten to spring of kindergarten, SLOPE 1: spring of kindergarten to spring of first grade, SLOPE 3: spring of first grade to spring of third grade), and predictors at each of the three levels was modeled. This model provided superior fit to the data as compared with a model including a single linear slope term (i.e., significantly smaller deviance statistic), and it provided greater interpretability. The Level 1 model represents how each child changes over time; the Level 2 model represents variation in growth parameters among children within the same school; the Level 3 model represents variation in growth parameters across schools. The following equations represent each of these models:

$$\text{Level 1: } Y_{tij} = \pi_{0ij} + \pi_{1ij}(\text{SLOPE K}) + \pi_{2ij}(\text{SLOPE 1}) \\ + \pi_{3ij}(\text{SLOPE 3}) \dots + \pi_{pij}(a_{pij}) + \epsilon_{ij},$$

where  $Y_{tij}$  represents child  $i$  in school  $j$ 's reading performance at time  $t$ ;  $\pi_{0ij}$  represents the intercept or the reading performance of child  $i$  in school  $j$  at the initial assessment;  $\pi_{1ij}$  represents the linear slope or the monthly learning rate in reading of child  $i$  in school  $j$  between the fall and the spring of kindergarten;  $\pi_{2ij}$  represents the monthly learning rate in reading of child  $i$  in school  $j$  between the spring of kindergarten and the spring of first grade;  $\pi_{3ij}$  represents the monthly learning rate in reading of child  $i$  in school  $j$  between the spring of first grade and the spring of third grade;  $\pi_{pij}$  represents the strength and the direction of association between predictor variables  $a_{pij}$  and child  $i$  in school  $j$ 's performance at time  $t$ ; and  $\epsilon_{ij}$  represents the time-specific error of child  $i$  in school  $j$  at time  $t$ .

$$\text{Level 2: } \pi_{0ij} = \beta_{00j} \dots + \beta_{0pj}(X_{0pj}) + r_{ij}, \\ \pi_{1ij} = \beta_{10j} \dots + \beta_{1pj}(X_{1pj}) + r_{ij}, \\ \pi_{2ij} = \beta_{20j} \dots + \beta_{2pj}(X_{2pj}) + r_{ij}, \text{ and} \\ \pi_{3ij} = \beta_{30j} \dots + \beta_{3pj}(X_{3pj}) + r_{ij},$$

where  $\beta_{00j}$  represents the mean intercept or the reading performance within school  $j$  at the initial assessment;  $\beta_{10j}$  represents the mean linear slope or the learning rate in reading within school  $j$  between the fall and the spring of kindergarten;  $\beta_{20j}$  represents the linear slope or the learning rate in reading within school  $j$  between the spring of kindergarten and the spring of first grade;  $\beta_{30j}$  represents the linear slope or the learning rate in reading within



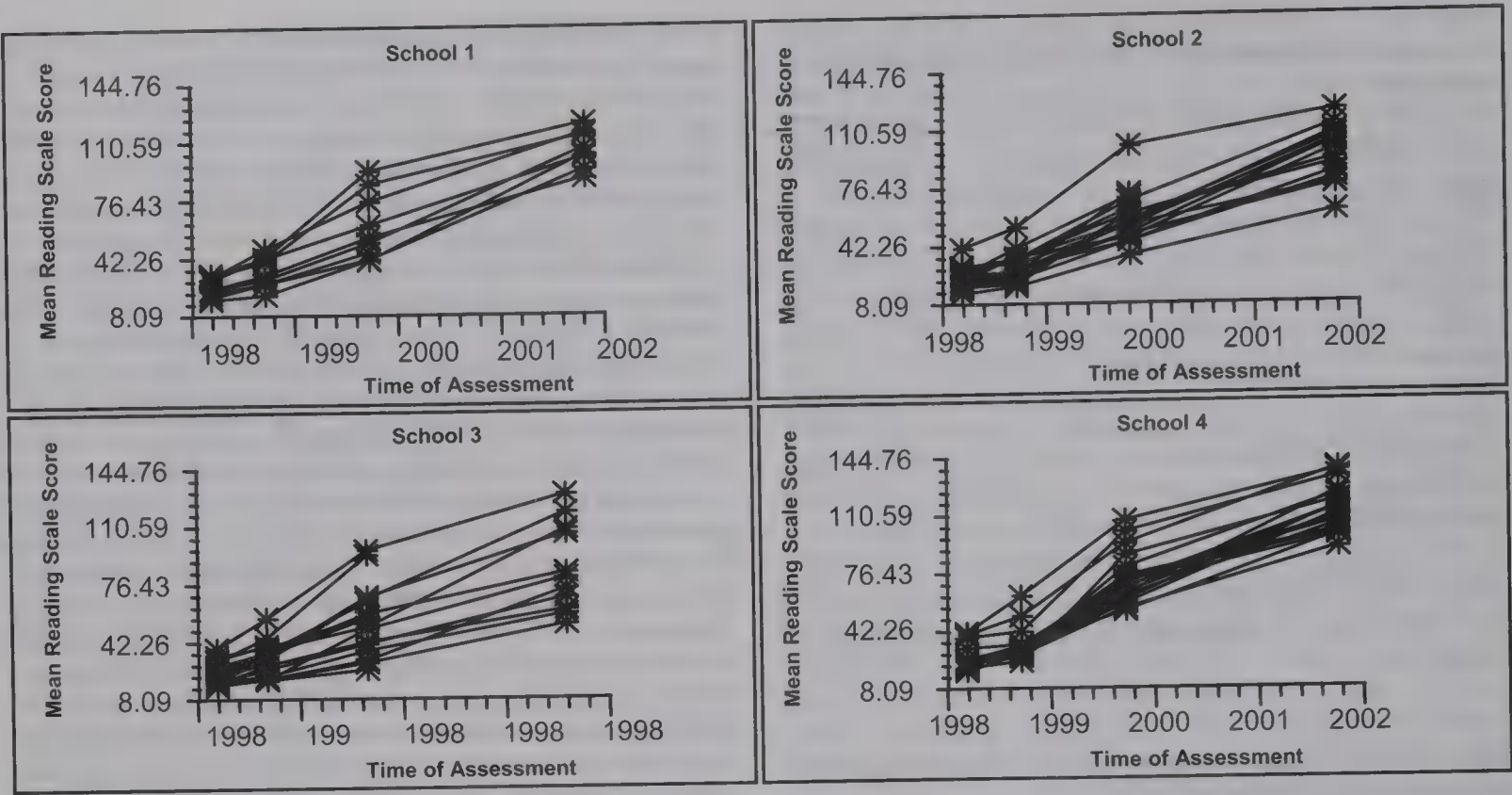


Figure 1. Smoothed plots of reading trajectories of a random subsample of children in four schools.

school  $j$  between the spring of first grade and the spring of third grade;  $\beta_{0pj}$  represents the strength and the direction of association between predictor variables  $X_{0pj}$  and child  $i$  in school  $j$ 's reading performance at the initial assessment,  $\pi_{0ij}$ ;  $\beta_{x_{pj}}$  represents the strength and the direction of association between predictor variables  $X_{x_{pj}}$  and child  $i$  in school  $j$ 's linear learning rate in reading during the three periods of interest,  $\pi_{1ij}$ ,  $\pi_{2ij}$ ,  $\pi_{3ij}$ ; and  $r_{ij}$  represents school  $j$ 's deviation from the reading score predicted by the model.

Level 3:  $\beta_{00j} = \gamma_{000} \dots + \gamma_{0p0}(X_{0p0}) + u_{00j}$ ,  
 $\beta_{10j} = \gamma_{100} \dots + \gamma_{1p0}(X_{1p0}) + u_{10j}$ ,  
 $\beta_{20j} = \gamma_{200} \dots + \gamma_{2p0}(X_{2p0}) + u_{20j}$ , and  
 $\beta_{30j} = \gamma_{300} \dots + \gamma_{3p0}(X_{3p0}) + u_{30j}$ ,

where  $\gamma_{000}$  represents the mean intercept or the reading performance at the initial assessment across schools;  $\gamma_{100}$  represents the mean linear slope or the learning rate in reading across schools between the fall and the spring of kindergarten;  $\gamma_{200}$  represents the mean learning rate in reading across schools between the spring of kindergarten and the spring of first grade;  $\gamma_{300}$  represents the mean learning rate in reading across schools between the spring of first grade and the spring of third grade;  $\gamma_{0p0}$  represents the strength and the direction of association between predictor variables  $X_{0p0}$  and mean initial reading performance within school  $j$ ,  $\beta_{00j}$ ;  $\gamma_{x_{p0}}$  represents the strength and the direction of association between predictor variables  $X_{x_{p0}}$  and mean linear learning rate within school  $j$  during the three periods of interest,  $\beta_{10j}$ ,  $\beta_{20j}$ ,  $\beta_{30j}$ ; and  $u_{ij}$  equals overall deviation across schools from the reading score predicted by the model.

All time varying and time invariant covariates were entered into the models per recommendations by Singer and Willett (2003). Time

varying variables include those that may change over time and that are measured at multiple time points, whereas time invariant predictors reflect data that do not change over time. For example, child's gender and race and child's attendance (or nonattendance) at center-based care the year prior to kindergarten are time invariant. Parents in the ECLS-K reported on aspects of the home environment across waves, making these variables time varying. Time-varying predictors predicted time-specific performance, whereas time-invariant predictors predicted the intercept (i.e., initial performance) and slope (i.e., growth over time) parameters. School-level variables, whether time varying or time invariant, predicted the intercept and slope parameters. Because time-varying predictors were entered at Level 2, they also predicted the intercept and slope parameters.

Because HLM cannot fit a four-level model, classroom and teacher characteristics were aggregated and used as covariates at the school level. Neighborhood variables that were associated with school environment were entered as covariates at Level 3. In HLM models, regression coefficients were interpreted as the value of the dependent variable when the contextual predictor and all the other variables in the model were equal to zero or the mean. Accordingly, with the exception of dummy codes, all contextual predictors were grand mean centered at zero to improve interpretability (Raudenbush & Bryk, 2002; Singer & Willet, 2003).

Results

Descriptive Statistics of Analytic Sample

To characterize family, neighborhood, and school characteristics of participating children, we present descriptive information for the analytic sample in kindergarten, first grade, and third grade in Table 1. As might be expected in a representative sample of U.S.

children, White children made up the majority of the sample (56%), and 44% of the children were non-White (16% were reported by parents as Black, 19% were reported as Hispanic; 4% were reported as Asian, and 5% were reported as other). English was the main language in about 9 of 10 homes, and about 1 in 5 families had household incomes that place them below the federal poverty line.

Table 2 presents children's mean reading achievement. Overall, children's reading achievement and reading achievement variability increased over time. Significant socioeconomic differences of 11.1 points were observed between the lowest and the highest quintiles in children's reading. IRT scores were observed in the fall of kindergarten. These differences increased over the first 4 years of schooling. The reading gap between the lowest and the highest quintiles grew by 2.8 points to 13.9 points from the fall to the spring of kindergarten. By the spring of first grade, the gap grew by 9.7 points to 23.6 points. By the end of third grade, it grew another 3.6 points to 27.2 points. This is an overall increase of 16.1 points in the reading score gap between the poorest children and the most affluent children from kindergarten to third grade, though most of the increase occurred in the first grade year. In the following section, we more fully discuss the meaning of these reading point differences.

### Reading Trajectories

Initially, an unconditional model of reading development was estimated to establish the functional form of reading growth in the full sample. Reading achievement was modeled as a linear function of children's initial reading (intercept), growth per month in reading between the fall and the spring of kindergarten (linear slope), monthly reading growth between the spring of kindergarten and the spring of first grade (linear slope), monthly reading growth between the spring of first grade and the spring of third grade (linear slope), and measurement error. The estimated unconditional model is represented by the following equations.

$$\text{Level 1: } \text{READ}_{ij} = \pi_{0ij} + \pi_{1ij}(\text{SLOPE K}) + \pi_{2ij}(\text{SLOPE 1}) \\ + \pi_{3ij}(\text{SLOPE 3}) + \varepsilon_{ij}.$$

$$\text{Level 2: } \pi_{0ij} = \beta_{00j} + r_{ij},$$

$$\pi_{1ij} = \beta_{10j},$$

$$\pi_{2ij} = \beta_{20j}, \text{ and}$$

$$\pi_{3ij} = \beta_{30j}.$$

$$\text{Level 3: } \beta_{00j} = \gamma_{000} + u_{00j},$$

$$\beta_{10j} = \gamma_{100},$$

$$\beta_{20j} = \gamma_{200}, \text{ and}$$

$$\beta_{30j} = \gamma_{300}.$$

Results from the unconditional model suggested that across children, mean initial reading achievement was 25.72 points ( $p < .001$ ), and children gained an average of 1.86 points per month in reading between the fall and the spring of kindergarten ( $p < .001$ ), 2.42 points per month in reading between the spring of kindergarten and the spring of first grade ( $p < .001$ ), and 1.66 points per

month in reading between the spring of first grade and the spring of third grade ( $p < .001$ ). Thus, the rate of children's reading growth differed across the three time periods, with the most rapid acquisition of skills taking place between the spring of kindergarten and the spring of first grade. See Figure 1 for examples of children's smoothed reading trajectories in a random subsample of four schools. The figure illustrates the spike or the heightened reading skill acquisition occurring between the spring of kindergarten and the spring of first grade. In addition, as noted, there was significant variation around intercept and linear slope terms, suggesting that reading trajectories varied across children.

Next, a conditional model was estimated with child demographic characteristics included as predictors of children's initial reading scores and monthly rates of growth in reading across the three time periods.<sup>1</sup>

$$\text{Level 1: } \text{READ}_{ij} = \pi_{0ij} + \pi_{1ij}(\text{SLOPE K}) + \pi_{2ij}(\text{SLOPE 1}) \\ + \pi_{3ij}(\text{SLOPE 3}) + \varepsilon_{ij}.$$

$$\text{Level 2: } \pi_{0ij} = \beta_{00j}(\text{AGE}) + \beta_{01j}(\text{FEMALE}) \\ + \beta_{02j}(\text{WHITE}) + \beta_{03j}(\text{SES}) + r_{ij},$$

$$\pi_{1ij} = \beta_{10j}(\text{AGE}) + \beta_{11j}(\text{FEMALE}) \\ + \beta_{12j}(\text{WHITE}) + \beta_{13j}(\text{SES}),$$

$$\pi_{2ij} = \beta_{20j}(\text{AGE}) + \beta_{21j}(\text{FEMALE}) \\ + \beta_{22j}(\text{WHITE}) + \beta_{23j}(\text{SES}), \text{ and}$$

$$\pi_{3ij} = \beta_{30j}(\text{AGE}) + \beta_{31j}(\text{FEMALE}) \\ + \beta_{32j}(\text{WHITE}) + \beta_{33j}(\text{SES}).$$

$$\text{Level 3: } \beta_{00j} = \gamma_{000} + u_{00j},$$

$$\beta_{01j} = \gamma_{010},$$

$$\beta_{02j} = \gamma_{020},$$

$$\beta_{03j} = \gamma_{030},$$

$$\beta_{10j} = \gamma_{100},$$

$$\beta_{11j} = \gamma_{110},$$

$$\beta_{12j} = \gamma_{120},$$

<sup>1</sup> In the equations, READ = child's reading IRT score; SUPPORT = neighborhood support for the school's goals and activities (as reported by school administrators); AGE = child age; FEMALE = child is female; WHITE = child is White; SES = child's socioeconomic status; PRE K = center-based care prior to kindergarten; HOMELIT = home literacy environment; BOOKS = number of books in the home; INVOLVE = parent involvement in the school; STRAIN = parental role strain; WARMTH = parental warmth; HPROBS = neighborhood problems; HSAFETY = neighborhood safety; BAD = bad conditions near the school; SUBURB = suburban neighborhood; SPROBS = school neighborhood problems; BELOW = number of children reading below grade; INSTRUCT = classroom instruction practices; EXPERIENCE = teacher's number of years experience teaching; PREPARATION = teacher's preparation to teach; POVERTY = school poverty concentration; and PRIVATE = private school.



Table 2  
Distributions of Variables Used in HLM Models by SES Quintiles

Variable	Quintile 1			Quintile 2			Quintile 3			Quintile 4			Quintile 5			Total		
	M	SE	%	M	SE	%	M	SE	%	M	SE	%	M	SE	%	M	SE	%
Reading fall K	21.8	0.31		24.7	0.28		26.7	0.23		28.8	0.26		33.1	0.37		27.5	0.22	
Reading spring K	31.7	0.51		35.7	0.45		38.2	0.31		40.5	0.36		45.6	0.49		39.1	0.30	
Reading spring 1st	55.3	0.74		63.5	0.64		67.7	0.52		72.2	0.66		78.9	0.62		69.3	0.48	
Reading spring 3rd	92.4	0.85		102.2	0.62		107.9	0.57		112.4	0.57		119.6	0.46		108.8	0.55	
Home literacy at K	6.2	0.05		6.5	0.04		6.7	0.05		7.0	0.04		7.2	0.03		6.7	0.02	
Home literacy at 1st	6.2	0.07		6.6	0.05		6.7	0.03		6.8	0.04		7.2	0.04		6.7	0.02	
Home literacy at 3rd	6.1	0.07		6.1	0.05		6.2	0.04		6.4	0.05		6.6	0.04		6.3	0.03	
Books in home in K	32.6	1.7		56.7	1.6		74.5	1.8		88.2	1.9		107.7	1.7		72.8	1.5	
Books in home in 1st	40.5	2.1		75.1	2.5		100.9	3.5		113.7	3.2		155.2	3.7		96.5	2.3	
Books in home in 3rd	57.1	4.7		95.9	4.3		111.0	4.0		140.8	6.0		171.2	4.2		115.8	3.1	
Involvement in school at K	2.5	0.07		3.3	0.06		3.7	0.05		4.1	0.04		4.4	0.04		3.6	0.04	
Involvement in school at 1st	2.8	0.06		3.6	0.05		4.1	0.05		4.4	0.05		4.8	0.03		3.9	0.04	
Involvement in school at 3rd	3.2	0.05		3.7	0.05		4.1	0.04		4.5	0.04		4.9	0.03		4.1	0.03	
Parent role strain at K	1.7	0.03		1.6	0.01		1.5	0.02		1.5	0.01		1.5	0.01		1.6	0.01	
Parent role strain at 1st	1.7	0.02		1.6	0.02		1.5	0.01		1.5	0.01		1.5	0.01		1.6	0.01	
Parent role strain at 3rd	1.7	0.03		1.6	0.03		1.5	0.02		1.5	0.02		1.5	0.02		1.6	0.01	
Parent warmth at K	3.7	0.01		3.7	0.01		3.7	0.01		3.7	0.01		3.7	0.01		3.7	0.01	
Parent warmth at 1st	3.7	0.01		3.7	0.01		3.7	0.01		3.7	0.01		3.7	0.01		3.7	0.01	
Parent warmth at 3rd	3.6	0.02		3.6	0.02		3.6	0.01		3.7	0.01		3.6	0.01		3.6	0.01	
Home neighborhood problems at K	0.81	0.07		0.55	0.04		0.39	0.03		0.31	0.03		0.19	0.01		0.37	0.03	
Home neighborhood problems at 1st	0.66	0.04		0.49	0.04		0.33	0.03		0.24	0.02		0.16	0.01		0.33	0.03	
Home neighborhood problems at 3rd	0.81	0.08		0.48	0.04		0.32	0.04		0.23	0.03		0.15	0.01		0.35	0.02	
Home neighborhood safety at K	2.4	0.03		2.6	0.02		2.7	0.02		2.8	0.02		2.8	0.01		2.7	0.01	
Home neighborhood safety at 1st	2.5	0.02		2.7	0.02		2.7	0.02		2.8	0.02		2.9	0.01		2.7	0.01	
Home neighborhood safety at 3rd	2.5	0.02		2.7	0.02		2.8	0.02		2.8	0.01		2.9	0.01		2.7	0.01	
School neighborhood problems at K	3.4	0.25		2.7	0.15		2.4	0.16		2.1	0.14		1.4	0.12		2.1	0.14	
School neighborhood problems at 1st	3.6	0.19		2.7	0.15		2.0	0.11		1.7	0.11		1.2	0.09		1.9	0.13	
School neighborhood problems at 3rd	3.4	0.19		2.6	0.13		2.0	0.13		1.4	0.08		1.1	0.08		1.8	0.12	
Community support for learning at K	3.9	0.05		4.0	0.04		4.0	0.04		4.2	0.04		4.3	0.05		4.1	0.03	
Community support for learning at 1st	3.9	0.08		4.1	0.05		4.1	0.04		4.2	0.04		4.4	0.04		4.1	0.04	
Community support for learning at 3rd	3.9	0.06		4.1	0.05		4.1	0.04		4.2	0.04		4.4	0.04		4.2	0.04	
Bad conditions near school at K	0.8	0.07		0.6	0.05		0.5	0.04		0.4	0.04		0.4	0.04		0.5	0.04	
Bad conditions near school at 1st	1.4	0.05		1.3	0.03		1.2	0.02		1.2	0.02		1.1	0.01		1.2	0.02	
Bad conditions near school at 3rd	1.1	0.05		0.9	0.04		0.8	0.03		0.8	0.03		0.7	0.02		0.9	0.03	
Peers reading below grade at K	4.7	0.19		3.8	0.13		3.3	0.11		3.1	0.10		2.5	0.11		3.3	0.11	
Peers reading below grade at 1st	5.8	0.24		4.9	0.15		4.4	0.15		4.0	0.13		3.6	0.11		4.5	0.15	
Peers reading below grade at 3rd	6.6	0.32		5.4	0.21		5.0	0.14		4.5	0.15		3.6	0.10		4.9	0.16	
Attends high-poverty school in K			67%			59%			48%			37%			26%			48%
Attends high-poverty school in 1st			60%			51%			42%			31%			22%			41%
Attends high-poverty school in 3rd			37%			24%			15%			9%			4%			18%
Literacy instruction at K	11.7	0.06		11.7	0.06		11.7	0.05		11.7	0.05		11.8	0.05		11.7	0.07	
Literacy instruction at 1st	18.1	0.18		17.9	0.15		17.7	0.15		17.5	0.18		17.0	0.19		17.6	0.12	
Literacy instruction at 3rd	23.4	0.21		23.3	0.19		23.1	0.24		23.3	0.20		23.3	0.21		23.3	0.14	
Teacher experience at K	17.4	0.64		18.9	0.81		19.2	0.72		18.9	0.67		18.3	0.64		18.5	0.53	
Teacher experience at 1st	14.8	0.65		16.6	0.61		17.2	0.59		16.6	0.54		17.2	0.59		17.2	0.54	
Teacher experience at 3rd	15.7	0.75		16.6	0.60		17.5	0.57		16.5	0.42		16.2	0.54		17.3	0.47	
Teacher preparation at K	18.3	0.28		18.3	0.25		18.6	0.23		18.8	0.22		18.9	0.27		18.5	0.23	
Teacher preparation at 1st	17.8	0.26		17.8	0.22		18.4	0.23		18.3	0.20		18.6	0.23		18.5	0.19	
Teacher preparation at 3rd	16.0	0.27		16.5	0.21		16.1	0.22		16.4	0.20		16.4	0.20		16.3	0.17	

Note. For Quintile 1,  $n = 765,187$ ; for Quintile 2,  $n = 767,500$ ; for Quintile 3,  $n = 783,296$ ; for Quintile 4,  $n = 758,962$ ; for Quintile 5,  $n = 758,962$ . Total  $N = 3,833,907$ . Numbers in parentheses are standard errors. HLM = hierarchical linear modeling; SES = socioeconomic status; K = kindergarten; 1st = first grade; 3rd = third grade.

$$\beta_{13j} = \gamma_{130},$$

$$\beta_{20j} = \gamma_{200},$$

$$\beta_{21j} = \gamma_{210},$$

$$\beta_{23j} = \gamma_{230},$$

$$\beta_{31j} = \gamma_{310},$$

$$\beta_{32j} = \gamma_{320}, \text{ and}$$

$$\beta_{33j} = \gamma_{330}.$$

According to this piecewise model, children who were older, female, and from higher SES households had higher reading scores at the initial assessment. In addition, female and high-SES children gained reading skills at a faster rate than did their peers between the fall and the spring of kindergarten; White, female, and high-SES children gained reading skills at a faster rate than did other children between the spring of kindergarten and the spring of first grade; and younger, White, and high-SES children gained reading skills at a faster rate between the spring of first grade and the spring of third grade. Of most importance are the parameter estimates representing the effect of family SES on children's initial reading skills and monthly growth during the three periods.

When other child characteristics were controlled for, as SES increased, so did children's reading achievement at the initial assessment in the fall of kindergarten. In fact, children at 1 standard deviation above the mean in SES had reading scores at the initial kindergarten assessment that were 2.24 points higher than children of average SES and 4.48 points higher than children 1 standard deviation below the mean in SES. Given that 1.86 was the average monthly learning rate for the full sample in the kindergarten year, a 2.24-point advantage in initial reading is a head start, for children 1 standard deviation above average in SES, of more than 1 month relative to children of average SES. It is a head start of more than 2 months relative to children 1 standard deviation below mean SES.

Moreover, on the basis of the model, we expected that children who differed by 1 standard deviation in SES would differ by 0.10 points per month in the rate of reading skill acquisition during the kindergarten year, 0.19 points per month between the spring of kindergarten and the spring of first grade, and 0.02 points per month between the spring of first grade and the spring of third grade. Thus, higher SES relates to higher initial achievement in the fall of kindergarten and more rapid reading growth per month across the three periods of interest. The rate of these advantages in

reading growth differs across the three time periods, with the period between the spring of kindergarten and the spring of first grade being that of greatest differentiation by SES (see Table 3 for parameter estimates).

### *Family Characteristics, Reading Achievement, and Growth*

Next, family covariates (home literacy environment, school involvement, parental role strain, number of books in home, and parental warmth) were added to the model to assess the extent to which SES gaps in initial reading and monthly reading growth rates during the three periods of interest could be explained by family characteristics beyond demographic characteristics. This model revealed the effects of family covariates on initial reading scores, monthly reading growth during the three periods, and time-specific reading performance beyond what could be predicted by children's underlying trajectories and demographic characteristics.

$$\begin{aligned} \text{Level 1: } \text{READ}_{tij} = & \pi_{0ij} + \pi_{1ij}(\text{SLOPE K}) + \pi_{2ij}(\text{SLOPE 1}) \\ & + \pi_{3ij}(\text{SLOPE 3}) + \pi_{4ij}(\text{HOMELIT}) + \pi_{5ij}(\text{BOOKS}) \\ & + \pi_{6ij}(\text{INVOLVE}) + \pi_{7ij}(\text{STRAIN}) + \pi_{8ij}(\text{WARMTH}) + \varepsilon_{ij}. \end{aligned}$$

$$\begin{aligned} \text{Level 2: } \pi_{0ij} = & \beta_{00j} + \beta_{01j}(\text{AGE}) + \beta_{02j}(\text{FEMALE}) \\ & + \beta_{03j}(\text{WHITE}) + \beta_{04j}(\text{SES}) \\ & + \beta_{05j}(\text{PRE K}) + r_{ij}, \\ \pi_{1ij} = & \beta_{10j} + \beta_{11j}(\text{AGE}) + \beta_{12j}(\text{FEMALE}) \\ & + \beta_{13j}(\text{WHITE}) + \beta_{14j}(\text{SES}) \\ & + \beta_{15j}(\text{PRE K}), \\ \pi_{2ij} = & \beta_{20j} + \beta_{21j}(\text{AGE}) + \beta_{22j}(\text{FEMALE}) \\ & + \beta_{23j}(\text{WHITE}) + \beta_{24j}(\text{SES}) + \beta_{25j}(\text{PRE K}), \\ \pi_{3ij} = & \beta_{30j} + \beta_{31j}(\text{AGE}) + \beta_{32j}(\text{FEMALE}) \\ & + \beta_{33j}(\text{WHITE}) + \beta_{34j}(\text{SES}) + \beta_{35j}(\text{PRE K}), \\ \pi_{4ij} = & \beta_{40j}, \\ \pi_{5ij} = & \beta_{50j}, \\ \pi_{6ij} = & \beta_{60j}, \end{aligned}$$

Table 3  
*Parameter Estimates and Standard Errors for the Child Model*

Independent variable	Intercept	K linear slope	K-1 linear slope	1-3 linear slope
Age at initial assessment	0.42 (0.04)***	-0.00 (0.00)	0.00 (0.00)	-0.01 (0.00)***
Female	1.78 (0.29)***	0.11 (0.03)**	0.15 (0.03)***	-0.00 (0.02)
White	0.60 (0.43)	0.06 (0.04)	0.14 (0.04)**	0.15 (0.02)***
SES	2.24 (0.15)***	0.097 (0.02)***	0.189 (0.02)***	0.019 (0.01)*

Note. Numbers in parentheses are robust standard errors. K = kindergarten; K-1 = kindergarten to first grade; 1-3 = first grade to third grade; SES = socioeconomic status.

\*  $p \leq .05$ . \*\*  $p \leq .01$ . \*\*\*  $p \leq .001$ .



$\pi_{7ij} = \beta_{70j}$ , and  
 $\pi_{8ij} = \beta_{80j}$ .

Level 3:  $\beta_{00j} = \gamma_{000} + u_{00j}$ ,

$\beta_{01j} = \gamma_{010}$ ,  
 $\beta_{02j} = \gamma_{020}$ ,  
 $\beta_{03j} = \gamma_{030}$ ,  
 $\beta_{04j} = \gamma_{040}$ ,  
 $\beta_{05j} = \gamma_{050}$ ,  
 $\beta_{10j} = \gamma_{100}$ ,  
 $\beta_{11j} = \gamma_{110}$ ,  
 $\beta_{12j} = \gamma_{120}$ ,  
 $\beta_{13j} = \gamma_{130}$ ,  
 $\beta_{14j} = \gamma_{140}$ ,  
 $\beta_{15j} = \gamma_{150}$ ,  
 $\beta_{20j} = \gamma_{200}$ ,  
 $\beta_{21j} = \gamma_{210}$ ,  
 $\beta_{22j} = \gamma_{220}$ ,  
 $\beta_{23j} = \gamma_{230}$ ,  
 $\beta_{24j} = \gamma_{240}$ ,  
 $\beta_{25j} = \gamma_{250}$ ,  
 $\beta_{30j} = \gamma_{300}$ ,  
 $\beta_{31j} = \gamma_{310}$ ,  
 $\beta_{32j} = \gamma_{320}$ ,  
 $\beta_{33j} = \gamma_{330}$ ,

$\beta_{34j} = \gamma_{340}$ ,  
 $\beta_{35j} = \gamma_{350}$ ,  
 $\beta_{40j} = \gamma_{400}$ ,  
 $\beta_{50j} = \gamma_{500}$ ,  
 $\beta_{60j} = \gamma_{600}$ ,  
 $\beta_{70j} = \gamma_{700}$ , and  
 $\beta_{80j} = \gamma_{800}$ .

Comparisons of the SES coefficient on the intercept and the three slope terms from this model with those of the previous demographic model indicated the extent to which mean differences in initial reading scores and learning rates in reading were attributable to family characteristics. In fact, with the addition of family variables, the SES-intercept coefficient was reduced by 16%, though the three slope parameters were not reduced. Thus, characteristics associated with the family context helped account for SES differences in children's initial reading scores, but they did not account for many of the SES differences in the rate at which children gained reading skills during the three periods of interest. Notably, children with richer home literacy environments ( $T = 6.20, p < .001$ , effect size [es] = .07), more books at home ( $T = 2.10, p < .05$ , es = .02), and parents who were less strained in the parenting role ( $T = -3.28, p < .001$ , es = .04) had enhanced time-specific reading performance during the first 4 years of school. On the basis of the model, children who attended center-based care the year prior to kindergarten were expected to have higher initial reading scores ( $T = 4.46, p < .001$ , es = .09), though receipt of such care was not expected to relate to reading growth during any of the three time periods through third grade (see Table 4).

*Neighborhood Characteristics, Reading Achievement, and Growth*

Next, neighborhood covariates were added to the model to allow assessment of the extent to which SES gaps in initial reading and

Table 4  
*Parameter Estimates and Standard Errors for the Family Model*

Independent variable	Intercept	K linear slope	K-1 linear slope	1-3 linear slope	Time-specific performance
Age at initial assessment	0.42 (0.04)***	-0.00 (0.00)	0.01 (0.00)	-0.01 (0.00)***	
Female	1.42 (0.29)***	0.10 (0.03)**	0.16 (0.04)***	-0.00 (0.02)	
White	0.15 (0.42)	0.05 (0.04)	0.13 (0.04)**	0.16 (0.02)***	
SES	1.88 (0.16)***	0.104 (0.03)***	0.188 (0.02)***	0.016 (0.01)	
Center-based care	1.49 (0.33)***	-0.07 (0.05)	-0.01 (0.04)	0.01 (0.02)	
Home literacy environment					0.47 (0.08)***
Books in home					0.00 (0.00)*
Involvement in school					0.17 (0.10)
Parental role strain					-1.25 (0.33)***
Parental warmth					0.31 (0.36)

Note. Numbers in parentheses are robust standard errors. K = kindergarten; K-1 = kindergarten to first grade; 1-3 = first grade to third grade; SES = socioeconomic status.  
\*  $p \leq .05$ . \*\*  $p \leq .01$ . \*\*\*  $p \leq .001$ .

monthly reading growth rates could be explained by neighborhood characteristics beyond children's demographic and family characteristics.

$$\begin{aligned} \text{Level 1: } \text{READ}_{ij} = & \pi_{0ij} + \pi_{1ij}(\text{SLOPE K}) + \pi_{2ij}(\text{SLOPE 1}) \\ & + \pi_{3ij}(\text{SLOPE 3}) + \pi_{4ij}(\text{HOMELIT}) + \pi_{5ij}(\text{BOOKS}) \\ & + \pi_{6ij}(\text{INVOLVE}) + \pi_{7ij}(\text{STRAIN}) + \pi_{8ij}(\text{WARMTH}) \\ & + \pi_{9ij}(\text{HPROBS}) + \pi_{10ij}(\text{HSAFETY}) + \epsilon_{ij}. \end{aligned}$$

$$\begin{aligned} \text{Level 2: } \pi_{0ij} = & \beta_{00j} + \beta_{01j}(\text{AGE}) + \beta_{02j}(\text{FEMALE}) \\ & + \beta_{03j}(\text{WHITE}) + \beta_{04j}(\text{SES}) + \beta_{05j}(\text{PRE K}) + r_{ij}, \\ \pi_{1ij} = & \beta_{10j} + \beta_{11j}(\text{AGE}) + \beta_{12j}(\text{FEMALE}) \\ & + \beta_{13j}(\text{WHITE}) + \beta_{14j}(\text{SES}) + \beta_{15j}(\text{PRE K}), \\ \pi_{2ij} = & \beta_{20j} + \beta_{21j}(\text{AGE}) + \beta_{22j}(\text{FEMALE}) \\ & + \beta_{23j}(\text{WHITE}) + \beta_{24j}(\text{SES}) + \beta_{25j}(\text{PRE K}), \\ \pi_{3ij} = & \beta_{30j} + \beta_{31j}(\text{AGE}) + \beta_{32j}(\text{FEMALE}) \\ & + \beta_{33j}(\text{WHITE}) + \beta_{34j}(\text{SES}) + \beta_{35j}(\text{PRE K}), \\ \pi_{4ij} = & \beta_{40j}, \\ \pi_{5ij} = & \beta_{50j}, \\ \pi_{6ij} = & \beta_{60j}, \\ \pi_{7ij} = & \beta_{70j}, \\ \pi_{8ij} = & \beta_{80j}, \\ \pi_{9ij} = & \beta_{90j}, \text{ and} \\ \pi_{10ij} = & \beta_{100j}. \end{aligned}$$

$$\begin{aligned} \text{Level 3: } \beta_{00j} = & \gamma_{000} + \gamma_{001}(\text{SUBURB}) + \gamma_{002}(\text{SUPPORT}) \\ & + \gamma_{003}(\text{SPROBS}) + \gamma_{004}(\text{BAD}) + u_{00j}, \\ \beta_{01j} = & \gamma_{010}, \\ \beta_{02j} = & \gamma_{020}, \\ \beta_{03j} = & \gamma_{030}, \\ \beta_{04j} = & \gamma_{040}, \\ \beta_{05j} = & \gamma_{050}, \\ \beta_{10j} = & \gamma_{100} + \gamma_{101}(\text{SUBURB}) + \gamma_{102}(\text{SUPPORT}) \\ & + \gamma_{103}(\text{SPROBS}) + \gamma_{104}(\text{BAD}), \\ \beta_{11j} = & \gamma_{110}, \\ \beta_{12j} = & \gamma_{120}, \\ \beta_{13j} = & \gamma_{130}, \\ \beta_{14j} = & \gamma_{140}, \\ \beta_{15j} = & \gamma_{150}, \end{aligned}$$

$$\begin{aligned} \beta_{20j} = & \gamma_{200} + \gamma_{201}(\text{SUBURB}) + \gamma_{202}(\text{SUPPORT}) \\ & + \gamma_{203}(\text{SPROBS}) + \gamma_{204}(\text{BAD}), \end{aligned}$$

$$\beta_{21j} = \gamma_{210},$$

$$\beta_{22j} = \gamma_{220},$$

$$\beta_{23j} = \gamma_{230},$$

$$\beta_{24j} = \gamma_{240},$$

$$\beta_{25j} = \gamma_{250},$$

$$\begin{aligned} \beta_{30j} = & \gamma_{300} + \gamma_{301}(\text{SUBURB}) + \gamma_{302}(\text{SUPPORT}) \\ & + \gamma_{303}(\text{SPROBS}) + \gamma_{304}(\text{BAD}), \end{aligned}$$

$$\beta_{31j} = \gamma_{310},$$

$$\beta_{32j} = \gamma_{320},$$

$$\beta_{33j} = \gamma_{330},$$

$$\beta_{34j} = \gamma_{340},$$

$$\beta_{35j} = \gamma_{350},$$

$$\beta_{40j} = \gamma_{400},$$

$$\beta_{50j} = \gamma_{500},$$

$$\beta_{60j} = \gamma_{600},$$

$$\beta_{70j} = \gamma_{700},$$

$$\beta_{80j} = \gamma_{800},$$

$$\beta_{90j} = \gamma_{900}, \text{ and}$$

$$\beta_{100j} = \gamma_{1000}.$$

With the addition of neighborhood variables, the SES-intercept coefficient was reduced by 2%, the slope parameter for the period between the spring of kindergarten and the spring of first grade was reduced by 1%, and the slope parameter for the period between the spring of first grade and the spring of third grade was reduced by 2%. The kindergarten slope parameter was unchanged with the addition of neighborhood covariates. Thus, this model suggested that beyond family and demographic characteristics, neighborhood conditions did not account much for expected differences in children's initial reading achievement in the fall of kindergarten, but they did account for expected differences in rates of children's monthly reading growth as children matured. In addition, schools with poor conditions in the surrounding area ( $T = -2.19$ ,  $p < .05$ ,  $es = .03$ ) had constrained reading growth during the kindergarten year and during the period between the spring of kindergarten and the spring of first grade ( $T = -2.07$ ,  $p < .05$ ,  $es = .02$ ). Community support for learning was related to higher school mean reading scores at the initial kindergarten assessment ( $T = 2.79$ ,  $p < .01$ ,  $es = .1890$ ; see Table 5).

#### *School Characteristics, Reading Achievement, and Growth*

To assess the extent to which SES gaps in initial reading and monthly reading growth rates could be explained by school char-



Table 5  
Parameter Estimates and Standard Errors for the Neighborhood Model

Independent variable	Intercept	K linear slope	K-1 linear slope	1-3 linear slope	Time-specific performance
Age at initial assessment	0.42 (0.04) <sup>***</sup>	-0.00 (0.00)	0.01 (0.00)	-0.01 (0.00) <sup>***</sup>	
Female	1.41 (0.29) <sup>***</sup>	0.10 (0.03) <sup>**</sup>	0.16 (0.04) <sup>***</sup>	-0.00 (0.02)	
White	0.60 (0.41)	0.05 (0.05)	0.07 (0.05)	0.15 (0.03) <sup>***</sup>	
SES	1.85 (0.15) <sup>***</sup>	0.105 (0.02) <sup>***</sup>	0.177 (0.02) <sup>***</sup>	0.016 (0.01)	0.48 (0.08) <sup>***</sup>
Home literacy environment					0.00 (0.00) <sup>*</sup>
Books in home					0.18 (0.10)
Involvement in school					-1.21 (0.33) <sup>***</sup>
Parental role strain					0.37 (0.37)
Parental warmth					
Center-based care	1.47 (0.33) <sup>***</sup>	-0.07 (0.05)	-0.02 (0.04)	0.01 (0.02)	-0.17 (0.18)
Home neighborhood problems					0.02 (0.25)
Home neighborhood safety					
Suburban neighborhood	-0.30 (0.61)	-0.09 (0.05)	0.02 (0.04)	0.00 (0.03)	
Community support for learning	0.99 (0.36) <sup>**</sup>	0.03 (0.03)	0.00 (0.03)	-0.01 (0.01)	
School neighborhood problems	-0.16 (0.16)	0.01 (0.01)	-0.02 (0.01)	-0.01 (0.01)	
Bad conditions near school	0.13 (0.39)	-0.08 (0.04) <sup>*</sup>	-0.07 (0.03) <sup>*</sup>	-0.02 (0.02)	

Note. Number of schools = 939. Numbers in parentheses are robust standard errors. K = kindergarten; K-1 = kindergarten to first grade; 1-3 = first grade to third grade; SES = socioeconomic status.

\*  $p \leq .05$ . \*\*  $p \leq .01$ . \*\*\*  $p \leq .001$ .

acteristics, beyond children's demographic and family characteristics, we added school covariates to the model.

Level 1:  $READ_{ij} = \pi_{0ij} + \pi_{1ij}(\text{SLOPE K}) + \pi_{2ij}(\text{SLOPE 1})$   
 $+ \pi_{3ij}(\text{SLOPE 3}) + \pi_{4ij}(\text{HOMELIT}) + \pi_{5ij}(\text{BOOKS})$   
 $+ \pi_{6ij}(\text{INVOLVE}) + \pi_{7ij}(\text{STRAIN}) + \pi_{8ij}(\text{WARMTH}) + \varepsilon_{ij}$

Level 2:  $\pi_{0ij} = \beta_{00j} + \beta_{01j}(\text{AGE}) + \beta_{02j}(\text{FEMALE})$   
 $+ \beta_{03j}(\text{WHITE}) + \beta_{04j}(\text{SES}) + \beta_{05j}(\text{PRE K}) + r_{ij}$   
 $\pi_{1ij} = \beta_{10j} + \beta_{11j}(\text{AGE}) + \beta_{12j}(\text{FEMALE})$   
 $+ \beta_{13j}(\text{WHITE}) + \beta_{14j}(\text{SES}) + \beta_{15j}(\text{PRE K}),$   
 $\pi_{2ij} = \beta_{20j} + \beta_{21j}(\text{AGE}) + \beta_{22j}(\text{FEMALE})$   
 $+ \beta_{23j}(\text{WHITE}) + \beta_{24j}(\text{SES}) + \beta_{25j}(\text{PRE K}),$   
 $\pi_{3ij} = \beta_{30j} + \beta_{31j}(\text{AGE}) + \beta_{32j}(\text{FEMALE})$   
 $+ \beta_{33j}(\text{WHITE}) + \beta_{34j}(\text{SES}) + \beta_{35j}(\text{PRE K}),$   
 $\pi_{4ij} = \beta_{40j},$   
 $\pi_{5ij} = \beta_{50j},$   
 $\pi_{6ij} = \beta_{60j},$   
 $\pi_{7ij} = \beta_{70j}, \text{ and}$   
 $\pi_{8ij} = \beta_{80j}.$

Level 3:  $\beta_{00j} = \gamma_{000} + \gamma_{001}(\text{BELOW}) + \gamma_{002}(\text{INSTRUCT})$   
 $+ \gamma_{003}(\text{EXPERIENCE})$   
 $+ \gamma_{004}(\text{PREPARATION}) + \gamma_{005}(\text{POVERTY})$

$+ \gamma_{006}(\text{PRIVATE}) + u_{00j},$

$\beta_{01j} = \gamma_{010},$

$\beta_{02j} = \gamma_{020},$

$\beta_{03j} = \gamma_{030},$

$\beta_{04j} = \gamma_{040},$

$\beta_{05j} = \gamma_{050},$

$\beta_{10j} = \gamma_{100} + \gamma_{101}(\text{BELOW}) + \gamma_{102}(\text{INSTRUCT})$

$+ \gamma_{103}(\text{EXPERIENCE}) + \gamma_{104}(\text{PREPARATION})$

$+ \gamma_{105}(\text{POVERTY}) + \gamma_{106}(\text{PRIVATE}),$

$\beta_{11j} = \gamma_{110},$

$\beta_{12j} = \gamma_{120},$

$\beta_{13j} = \gamma_{130},$

$\beta_{14j} = \gamma_{140},$

$\beta_{15j} = \gamma_{150},$

$\beta_{20j} = \gamma_{200} + \gamma_{201}(\text{BELOW}) + \gamma_{202}(\text{INSTRUCT})$

$+ \gamma_{203}(\text{EXPERIENCE}) + \gamma_{204}(\text{PREPARATION})$

$+ \gamma_{205}(\text{POVERTY}) + \gamma_{206}(\text{PRIVATE}),$

$\beta_{21j} = \gamma_{210},$

$\beta_{22j} = \gamma_{220},$

$$\begin{aligned}
\beta_{23j} &= \gamma_{230}, \\
\beta_{24j} &= \gamma_{240}, \\
\beta_{25j} &= \gamma_{250}, \\
\beta_{30j} &= \gamma_{300} + \gamma_{301}(\text{BELOW}) + \gamma_{302}(\text{INSTRUCT}) \\
&+ \gamma_{303}(\text{EXPERIENCE}) + \gamma_{304}(\text{PREPARATION}) \\
&+ \gamma_{305}(\text{POVERTY}) + \gamma_{306}(\text{PRIVATE}), \\
\beta_{31j} &= \gamma_{310}, \\
\beta_{32j} &= \gamma_{320}, \\
\beta_{33j} &= \gamma_{330}, \\
\beta_{34j} &= \gamma_{340}, \\
\beta_{35j} &= \gamma_{350}, \\
\beta_{40j} &= \gamma_{400}, \\
\beta_{50j} &= \gamma_{500}, \\
\beta_{60j} &= \gamma_{600}, \\
\beta_{70j} &= \gamma_{700}, \text{ and} \\
\beta_{80j} &= \gamma_{800}.
\end{aligned}$$

With the addition of school variables, the SES-intercept coefficient was reduced by less than 1%, and the kindergarten slope parameter was reduced by 4%. The slope parameter for the period between the spring of kindergarten and the spring of first grade was reduced by 13%. Notably, the slope parameter for the period between the spring of first grade and the spring of third grade was

not reduced with school variables added. This suggested that although characteristics associated with the school environment accounted for a small portion of SES gaps in children's initial reading skills, these characteristics accounted for a larger portion of differences in the rates of children's monthly reading growth during the periods of interest, with the largest effect between the spring of kindergarten and the spring of first grade. Notably, in private schools ( $T = 1.92$ ,  $p < .05$ ,  $es = .13$ ), children had higher initial reading scores, and in schools with more children reading below grade level in classrooms ( $T = -2.94$ ,  $p < .01$ ,  $es = .19$ ), children had constrained reading performance at the initial assessment in the fall of kindergarten. In these schools, children evidenced slower average reading growth in the period between the spring of kindergarten and the spring of first grade ( $T = -2.25$ ,  $p < .05$ ,  $es = .03$ ). Children in high poverty schools evidenced slower reading growth between the spring of kindergarten and the spring of first grade ( $T = -2.04$ ,  $p < .05$ ,  $es = .02$ ). There was no evidence that teacher experience, preparation, or classroom literacy instruction influenced initial reading scores or monthly reading growth rates beyond children's anticipated trajectories and demographic and family characteristics (see Table 6). In other work at prekindergarten, elementary, and secondary grade levels, researchers have found no relation (see for example, Early et al., 2006), not even a negative relation (Goldhaber & Brewer, 1997), between teacher background and student achievement.

In sum, these models suggest that the family context best accounts for SES disparities in children's initial reading achievement as they enter school. In addition, schools and neighborhoods best explain SES gaps in children's rates of monthly reading growth, particularly as children mature, whereas the families appear to do less to account for such gaps. The period between the spring of kindergarten and the spring of first grade is the one of the most rapid reading growth for children and the greatest differentiation on demographic, school, and neighborhood characteristics for children. It is important to note, however, that much of the SES gap in

Table 6  
Parameter Estimates and Standard Errors for the School Model

Independent variables	Intercept	K linear slope	K-1 linear slope	1-3 linear slope	Time-specific performance
Age at initial assessment	0.43 (0.04)***	-0.00 (0.00)	0.01 (0.00)	-0.01 (0.00)***	
Female	1.47 (0.29)***	0.10 (0.03)**	0.16 (0.03)***	-0.00 (0.02)	
White	0.55 (0.41)	0.04 (0.05)	0.09 (0.05)*	0.16 (0.03)***	
SES	1.88 (0.16)***	0.101 (0.02)**	0.163 (0.02)***	0.019 (0.01)*	
Home literacy environment					0.47 (0.08)***
Books in home					0.00 (0.00)*
Involvement in school					0.17 (0.10)
Parental role strain					-1.21 (0.33)***
Parental warmth					0.37 (0.37)
Center-based care	1.44 (0.33)***	-0.07 (0.05)	-0.02 (0.04)	0.01 (0.02)	
Poor readers in classroom	-0.44 (0.15)**	-0.02 (0.01)	-0.02 (0.01)*	0.00 (0.01)	
Literacy instruction	0.06 (0.05)	-0.00 (0.00)	0.00 (0.00)	-0.00 (0.00)	
Teacher experience	-0.03 (0.03)	-0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	
Teacher preparation	-0.03 (0.11)	-0.00 (0.01)	0.01 (0.01)	-0.00 (0.00)	
High-poverty school	0.03 (0.66)	0.02 (0.05)	-0.09 (0.05)*	-0.02 (0.03)	
Private school	1.64 (0.85)*	-0.05 (0.08)	0.07 (0.06)	-0.06 (0.04)	

Note. Number of schools = 939. Numbers in parentheses are robust standard errors. K = kindergarten; K-1 = kindergarten to first grade; 1-3 = first grade to third grade; SES = socioeconomic status.

\*  $p \leq .05$ . \*\*  $p \leq .01$ . \*\*\*  $p \leq .001$ .



initial reading achievement and in monthly learning rates remains unexplained by the variables included in the model. Table 7 includes random effects for all models.

Discussion

In the present study, we investigated the reading trajectories of children of differing socioeconomic backgrounds, from kindergarten to third grade, using a nationally representative sample. In addition, a special concern was understanding the unique and cumulative role of family, neighborhood, and school contexts in accounting for disparities in reading achievement during this time period. First, when considering family context, the following were all significantly related to reading outcomes: home literacy environment, number of books owned, parent distress, and receipt of center-based care. This result suggests that both family resource and/or investment models (e.g., books in home; Yeung, Linver, & Brooks-Gunn, 2002) and family process models (e.g., parent distress; Yeung et al., 2002) help explain reading outcomes. Of notable importance to reading outcomes are social relationships and processes (e.g., parental role strain).

Next, each of the contexts was associated with SES gaps in initial reading, reading achievement at specific points in time, and/or monthly learning rates. However, resources, experiences, and relationships associated with the family context were most closely associated with reading gaps at the initial kindergarten assessment. The importance of the experiences within the family context for early reading is hardly surprising because most children had experienced little schooling by the first assessment. Specifically, the results suggest that the relation between SES and children's initial reading competence is mediated by home literacy environment, number of books available within the home to the child, parental involvement in the school, parental role strain and warmth, and provision of center-based care prior to kindergarten. This is a useful finding. Although it may be difficult to alter family SES quickly enough to make a difference for young children, interventions can reduce the adverse effects of family mediators. The results of this study underscore the wisdom of programs that direct resources to strengthening family literacy environment, encouraging parental involvement in schools, and reducing parental role strain.

Although family factors were most strongly associated with SES differences in children's reading competence at the initial assessment, they were less associated with differences in gains or lack of progress in children's reading competence up to third grade. Conversely, experiences and resources associated with school and neighborhood conditions were understandably less associated with SES differences in initial reading scores but more closely associated with differences in children's reading skills acquisition rate, particularly during the period of most rapid reading growth—between the spring of kindergarten and the spring of first grade. Thus, there is a reversal in the relative relation of these settings once children enter school, such that qualities of school and community are associated with differences in reading development to a greater extent than is family life. Accordingly, family life can be viewed as being strongly associated with the starting point of children's reading competence, with other ecological settings being more strongly associated with children's reading progress. It is important to note that even in the presence of family characteristics, neighborhood and school characteristics were associated with children's reading performance. This reinforces the need to focus on characteristics and qualities of family life in preschool children and to widen the emphasis to include schools and neighborhoods as children enter school. It is important to note that these findings are consistent with the theoretical framework guiding the study—specifically that the relative salience of contexts may shift over time. As Bronfenbrenner (1989) noted, not only does the child change over time but the systems with which the child interacts change as well. Consequently, the nature of the child's interactions with these systems and their effect on the child may change over time. This is consistent with work by others (Aber, Gephart, Brooks-Gunn, & Connell, 1997) that recognized the differential effects of environmental contexts across developmental epochs.

Few studies, if any, have examined the relation between neighborhood context and reading development of the very young. This study provides preliminary evidence that neighborhood context may in fact be related to children's growth in reading. Specifically, community support for the school and poor physical conditions surrounding the school were associ-

Table 7  
*Random Effects for All Models*

Random effect	Child model			Family model			Neighborhood model			School model		
	Variance	$\chi^2$	df	Variance	$\chi^2$	df	Variance	$\chi^2$	df	Variance	$\chi^2$	df
Levels 1 and 2 variance components intercept ( $r_0$ )	84.77	135,806.99	8017	81.91	47,537.35***	7901	82.20	157,332.28***	7889	82.00	90,094.94	7901
Level 3 variance components intercept ( $u_{00}$ )	35.52	3,184.22***	936	35.50	3,170.02***	925	32.55	3,057.63***	920	31.52	2,976.21***	919
Deviance	216,069.11 (23)			207,251.43 (32)			206,276.11 (50)			207,092.46 (56)		

Note. In the last row, the numbers in parentheses are parameters. For the chi squares,  $N = df + 1$ .  
\*\*\*  $p \leq .001$ .

ated with children's reading. Future studies should explore in greater detail the nature and the role of neighborhood context on cognitive outcomes and on young children. Exploring the role of various schooling characteristics to reading outcomes and SES disparities in reading is relatively novel. Peers were a critical component of school context associated with children's reading outcomes in the present study. The number of children reading below grade and the presence of low-income peers were consistently associated with initial achievement and growth rates. The present findings, therefore, speak to the importance of grouping and segregation that occurs within schools. Notably, they are consistent with other work (e.g., Coleman, 1966) that suggested that "the social composition of the student body is more highly related to achievement, independent of the student's own social background, than is any school factor" (Coleman, 1966, p. 325). Thus, those designing educational policy to eliminate reading disparities must be aware of the relation of the presence and the concentration of peers with limited skills or fewer economic resources to students' skills and achievement. Teachers' experience, preparation, and classroom literacy instruction were not consistently related to children's reading outcomes beyond their expected trajectories. However, previous findings examining teacher effects have been mixed (Early et al. 2006). With a few notable exceptions, this work has not focused on the reading achievement of very young children. Taken together, the current findings highlight the ecological and systemic nature of development and the fact that resources and experiences across multiple contexts are associated with children's reading development. Moreover, these findings underscore the influence of SES. Consistent with the study's theoretical framework, in which developmental outcomes are a function of dynamic settings and contexts that interact with one another (Bronfenbrenner, 1979, 1989; Magnusson & Cairns, 1996), contexts produce unique and cumulative influences on reading outcomes and disparities.

In addition to describing the relation between family, school, and neighborhood characteristics to SES reading, our analyses also uncovered important phenomena in children's development of early reading competence. The SES-related differences in reading were stark at the initial kindergarten assessment following children's entry to school. Furthermore, the SES gap in reading between the poorest and the most affluent children grew as children progressed through school. The first grade year appears to be a critical time, as the SES gap in reading achievement grew the widest during this period. This may be due to the coming together of phonological processing, rapid accurate decoding, and comprehension by skilled readers during this time while the less skilled readers are still struggling with letter-sound combinations. It is important to note that this period was associated with the greatest differentiation across children on the basis of demographic characteristics and other ecological factors. For example, high-SES children started off as better readers and had more rapid progress than did low-SES children, across the three time periods, with the gap growing the most between the spring of kindergarten and the spring of first grade.

Our analyses also suggest a compounding effect of low quality environments. Children from low-SES homes grow up in home environments poor in literacy experiences. However, their disad-

vantage does not stop there. These children will often enter schools that have a higher proportion of poor children and those with low reading skills (Lee & Burkam, 2002; Phillips & Chin, 2004). Each of these factors is associated with poorer reading outcomes (Xue & Meisels, 2004). Schools may be overwhelmed in trying to serve concentrations of children who require considerable attention and resources. The present analyses support the policy that some schools have adopted of dispersing low-SES children across multiple schools so that there is not a high concentration of low-income children in a single school. Because school attendance areas correspond to neighborhoods and neighborhoods are often economically segregated, this policy often works against the cherished notion of neighborhood schools. As a consequence, stiff resistance to busing has arisen from parents who are unwilling to pay the price of dislocation necessary to implement these programs.

### *Limitations*

Several limitations mark the present study. First, a considerable portion of the SES gap remains unexplained by the models presented in the current analyses. Across models, we were able to account for only 17% of the initial SES gap and up to 16% of the growth rate gaps. Related to this, the selection of covariates for the growth models was not exhaustive of the ECLS-K dataset or of the set of contextual variables that may be related to reading achievement or SES. Certainly, there are other factors that may be linked to reading achievement and SES that were not explored in the analyses presented here. The inability to explain a larger proportion of the SES gap may be a result of the variables selected; that is, other variables may better capture contributing factors to the reading achievement gap. This difficulty also may be due to the limited measurement and the inherent weaknesses of relying on self-report measures of family practices and environmental conditions. Self-reports may be biased and affected by the social desirability of respondents. Others have similarly noted that parents with different educational backgrounds might vary in their ability to provide accurate reports (Dickinson & DeTemple, 1998). In addition, observational techniques or qualitative methods would be especially beneficial in capturing the true nature of children's contexts. For example, important aspects of teacher-child and parent-child interactions and social characteristics of those environments cannot be fully captured by a large-scale dataset like the ECLS-K. The researchers recognize that the advantages associated with using a large, nationally representative sample are tempered by the preclusion of other techniques that offer important insights in the contexts affecting children.

Finally, perhaps a more fruitful analysis of school versus nonschool factors would split gains into seasons and examine school learning versus summer learning. That is, by examining students' reading gains across years, as is done in the present analyses, it is more difficult to truly distinguish school effects from nonschool ones. We also may have underestimated school effects by not partitioning gains by seasons. Analyses by Downey, von Hippel, and Broh (2004) suggested that learning during the summer is more variable than is learning during the school year and that summer experiences are more responsible for the achievement gap. Others (see Rothstein, 2004) have made similar claims that children's summertime experiences



(and afterschool hours) play a role in contributing to disparities in achievement. In similar work, Burkam, Ready, Lee, and LoGerfo (2004) found that social stratification occurs during the summer months, with higher SES children learning more during these periods. However, highly interpretable models in the current analyses mimic children's overall reading and growth experiences, as losses and gains that children experience are cumulative across seasons.

### Strengths

The current study has several strengths. First, the use of a nationally representative sample of children eligible to attend kindergarten in the 1998–1999 school year greatly benefited the present analyses. Unlike previous studies in which convenience samples were used or only poor children were focused on, the present data provide a portrait of children from varying socioeconomic backgrounds and from across the nation. The data are not limited to low-income children living in urban environments, as often occurs in research on the effects of economic disadvantage. Such characteristics add to the external validity of this research. In addition, the current study benefits from the longitudinal nature of the ECLS-K. Rather than traditional cross-sectional designs, which often do not recognize the volatility in family characteristics linked to disadvantage and adversity, the current research may provide a more accurate portrait of the influence of the family environment on reading achievement over time. Moreover, the present study benefits from the comprehensive nature of the ECLS-K data on family and school contexts.

Finally, the current findings provide several implications for policy. To reduce the reading gap, one cannot place emphasis solely on improving experiences, resources, and interactions in any single context. The disadvantages that low-SES children face across contexts are great, and all contexts are associated with disparities in early reading. However, the present models suggest that families and characteristics associated with the home environment are most closely associated with SES gaps in children's reading achievement on entering school. Accordingly, efforts to improve children's home and family experiences, particularly before children arrive at kindergarten, may prove most important in addressing early disparities in reading achievement. For example, it may be beneficial to provide support services to reduce the effects of role strain unduly experienced by low-SES parents before children enter formal schooling. In addition, policies that increase parents' abilities to provide books in the home and to improve the home literacy environment for their children may also benefit young children. Similarly, the provision of prekindergarten childcare may prepare children for the school environment and provide foundational skills.

Researchers and policymakers should take care not to locate the source of reading disparities in children and families without also recognizing the constraints on families or the other contexts that may be at work. Like children themselves, family functioning exists within a larger context (Bronfenbrenner, 1986) so that parent behaviors, practices, and resources are a reflection of the experiences parents face outside of the home. For example, parents' behaviors and beliefs are not independent systems separate from the outside world; instead, they are

shaped by social, political, and economic forces occurring outside of the home. Policies and interventions that fail to recognize such intricacies may not ultimately succeed. Any policy seeking to improve conditions within the family system should seek to produce change in the contexts that families operate within as well.

### Conclusion

The present research suggests that no one solution or effort targeted to any single context will ameliorate the reading achievement gap. Ultimately, those involved in policy and intervention must recognize the ecological, dynamic nature of development and functioning. Children's development is multidetermined and embedded in dynamic, interconnected systems. As Farmer and Farmer (2001) contended, improvement due to interventions is likely to be short-lived if those involved in the interventions fail to understand the interconnection among systems and the ways in which multiple risks constrain developmental trajectories. Thus, as Rothstein (2004) argued, the answer to disparities in reading achievement lies across systems and in the amelioration of the disparate conditions that children and their families experience there.

### References

- Aber, J. L., Gephart, M. A., Brooks-Gunn, J., & Connell, J. P. (1997). Development in context: Implications for studying neighborhood effects. In J. Brooks-Gunn, G. J. Duncan, & J. L. Aber (Eds.), *Neighborhood poverty: Vol. 1. Context and consequences for children* (pp. 44–61). New York: Russell Sage Foundation.
- American Institutes of Research, & Cohen, J. (2005). AM Statistical Software (Version 0.06.03) [Computer software]. Available from American Institutes for Research: <http://am.air.org/default.asp>
- Arnold, D. H., & Doctoroff, G. L. (2003). The early education of socioeconomically disadvantaged children. *Annual Review of Psychology*, 54, 517–545.
- Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Cambridge, MA: Harvard University Press.
- Bronfenbrenner, U. (1986). Ecology of the family as a context for human development: Research perspectives. *Developmental Psychology*, 22, 723–742.
- Bronfenbrenner, U. (1989). Ecological systems theory. In R. Vasta (Ed.), *Annals of child development: Six theories of child development: Revised formulations and current issues* (pp. 1–103). Greenwich, CT: JAI Press.
- Bryant, D. M., Burchinal, M., Lau, L. B., & Sparling, J. J. (1994). Family and classroom correlates of Head Start children's developmental outcomes. *Early Childhood Research Quarterly*, 9, 289–309.
- Burkam, D. T., Ready, D. D., Lee, V. E., & LoGerfo, L. F. (2004). Social-class differences in summer learning between kindergarten and first grade: Model specification and estimation. *Sociology of Education*, 77, 1–31.
- Cairns, R. B., Elder, G. H., & Costello, E. J. (1996). *Developmental science*. New York: Cambridge University Press.
- Coleman, J. S. (1966). *Equality of educational opportunity* (ICSPR Study No. 6389). Washington DC: U.S. Government Printing Office.
- Dickinson, D. K. (2001). Book reading in preschool classrooms: Is recommended practice common? In D. K. Dickinson & P. O. Tabors (Eds.), *Beginning literacy with language: Young children learning at home and school* (pp. 175–204). Baltimore: Brookes Publishing.
- Dickinson, D. K., & DeTemple, J. (1998). Putting parents in the picture: Maternal reports of preschool literacy as a prediction of early reading. *Early Childhood Research Quarterly*, 13, 241–261.

- Downey, D. B., von Hippel, P. T., & Broh, B. A. (2004). Are schools the great equalizer? Cognitive inequality during the summer months and the school year. *American Sociological Review*, 69, 613–635.
- Duncan, S. E., & DeAvila, E. (1986). *Preschool Language Assessment Survey 2000 examiner's manual*. Monterey, CA: CTB/McGraw-Hill.
- Duncan, G. J., Yeung, W. J., Brooks-Gunn, J., & Smith, J. R. (1998). How much does childhood poverty affect the life chances of children? *American Sociological Review*, 63, 406–423.
- Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test, revised edition*. Circle Pines, MN: American Guidance Services.
- Early, D. M., Bryant, D. M., Pianta, R. C., Clifford, R. M., Burchinal, M. R., Ritchie, S., et al. (2006). Are teachers' education, major, and credentials related to classroom quality and children's academic gains in pre-kindergarten? *Early Childhood Research Quarterly*, 21, 174–195.
- Entwisle, D. R., & Alexander, K. L. (1993). Entry into school: The beginning school transition and educational stratification in the United States. *Annual Review of Sociology*, 19, 401–423.
- Evans, G. W. (2004). The environment of childhood poverty. *American Psychologist*, 59, 77–92.
- Farmer, T. W., & Farmer, E. M. Z. (2001). Developmental science, systems of care, and prevention of emotional and behavioral problems in youth. *American Journal of Orthopsychiatry*, 71, 171–181.
- Federal Interagency Forum on Child and Family Statistics. (2005). *America's children: Key indicators of well-being, 2005*. Washington, DC: U.S. Government Printing Office.
- Goldenberg, C. (2002). Making schools work for low-income families in the 21st century. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (pp. 211–231). New York: Guilford Press.
- Goldhaber, D. D., & Brewer, D. J. (1997). Why don't schools and teachers seem to matter? Assessing the impact of unobservables on educational productivity. *Journal of Human Resources*, 32, 505–523.
- Hart, B., & Risley, T. (1995). *Meaningful differences in the everyday experiences of young American children*. Baltimore: Brookes Publishing.
- Hess, R. D., Holloway, S. D., Dickson, W. P., & Price, G. G. (1984). Maternal variables as predictors of children's school readiness and later achievement in vocabulary and mathematics in sixth grade. *Child Development*, 55, 1902–1912.
- Huttenlocher, J., & Levine, S. C. (1990). *The Primary Test of Cognitive Skills (PTCS)*. New York: CTB/McGraw Hill.
- Kozol, J. (1991). *Savage inequalities: Children in America's schools*. New York: Harper Perennial.
- Lee, V. E., & Burkam, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children enter school*. Washington, DC: Economic Policy Institute.
- Leventhal, T., & Brooks-Gunn, J. (2000). The neighborhoods they live in: The effects of neighborhood residence on child and adolescent outcomes. *Psychological Bulletin*, 126, 309–337.
- Magnusson, D., & Cairns, R. B. (1996). Developmental science: Toward a unified framework. In R. B. Cairns, G. H. Elder, & E. J. Costello (Eds.), *Developmental science* (pp. 7–30). New York: Cambridge University Press.
- Markwardt, Jr., F. C., (1989). *Peabody Individual Achievement Test-Revised (PIAT-R)*. Circle Pines, MN: American Guidance Services.
- McLoyd, V. C. (1998). Socioeconomic disadvantage and child development. *American Psychologist*, 53, 185–204.
- Murray, A. D., & Hornbaker, A. V. (1997). Maternal directive and facilitative interaction styles: Associations with language and cognitive development of low risk and high risk toddlers. *Development and Psychopathology*, 9, 507–516.
- National Center for Education Statistics. (2004). ECLS-K longitudinal kindergarten-third grade public-use child file, CD-ROM and users' manual (NCES 2004–089). Washington, DC: U.S. Department of Education.
- National Research Council. (2000). *From neurons to neighborhoods: The science of early childhood development*. Washington, DC: National Academy Press.
- NICHD Early Child Care Research Network. (1998). Early child care and self-control, compliance, and problem behavior at 24 and 36 months. *Child Development*, 69, 1145–1170.
- NICHD Early Child Care Research Network. (2000). The relation of child care to cognitive and language development. *Child Development*, 71, 960–980.
- NICHD Early Child Care Research Network. (2004). Multiple pathways to early academic achievement. *Harvard Educational Review*, 74, 1–29.
- Phillips, M., & Chin, T. (2004). *School inequality: What do we know?* In K. Neckerman (Ed.), *Social inequality*. New York: Russell Sage Foundation.
- Pianta, R. C., LaParo, K. M., Payne, C. C., Cox, M. J., & Bradley, R. H. (2002). The relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *Elementary School Journal*, 102, 225–238.
- Pianta, R. C., & Walsh, D. (1996). *High-risk children in the schools: Creating sustaining relationships*. New York: Routledge.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2005). HLM 6 (Version 6.02) [Computer software]. Lincolnwood, IL: Scientific Software International.
- Raviv, T., Kessenich, M., & Morrison, F. J. (2004). A mediational model of the association between socioeconomic status and three-year old language abilities: The role of parenting factors. *Early Childhood Research Quarterly*, 19, 528–547.
- Reid, D. K., Hresko, W. P., & Hammill, D. D. (1981). *Test of Early Reading Ability (TERA-2)*. Austin, TX: PRO-ED.
- Rothstein, R. (2004). *Class and schools: Using social, economic, and educational reform to close the Black-White achievement gap*. Washington, DC: Economic Policy Institute.
- Singer, J., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford, England: University Press.
- Snow, C. E., Burns, S. M., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.
- Vernon-Feagans, L., Hammer, C. S., Miccio, A., & Manlove, E. (2002). Early language and literacy skills in low-income African American and Hispanic children. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (pp. 192–210). New York: Guilford Press.
- Weigel, D. J., Martin, S. S., & Bennett, K. K. (2005). Ecological influences of the home and the child-care center on preschool-age children's literacy development. *Reading Research Quarterly*, 40, 205–233.
- Whitehurst, G. J., & Lonigan, C. J. (1998). Child development and emergent literacy. *Child Development*, 69, 848–872.
- Woodcock, R. W., & Bonner, M. (1989). *Woodcock-Johnson III Tests of Achievement-Revised*. Itasca, IL: Riverside.
- Xue, Y., & Meisels, S. J. (2004). Early literacy instruction and learning to read in kindergarten. *American Educational Research Journal*, 41, 191–229.
- Yeung, W. J., Linver, M. R., & Brooks-Gunn, J. (2002). How money matters for young children's development: Parental investment and family processes. *Child Development*, 73, 1861–1879.

Received August 24, 2006

Revision received August 28, 2007

Accepted September 24, 2007 ■



# Preschool Home Literacy Practices and Children's Literacy Development: A Longitudinal Analysis

Michelle Hood, Elizabeth Conlon, and Glenda Andrews  
Griffith University

In this 3-year longitudinal study, the authors tested and extended M. Sénéchal and J. Le Fevre's (2002) model of the relationships between preschool home literacy practices and children's literacy and language development. Parent-child reading (Home Literacy Environment Questionnaire plus a children's Title Recognition Test) and parental teaching of letters, words, and name writing were assessed 6 months prior to children's school entry. The 143 children (55% male participants; mean age = 5.36 years,  $SD = 0.29$ ) attended Gold Coast, Australia government preschools. Parent-child reading and literacy teaching were only weakly correlated ( $r = .18$ ) and were related to different outcomes consistent with the original model. Age, gender, memory, and nonverbal ability were controlled. Parental teaching was independently related to R. W. Woodcock's (1997) preschool Letter-Word Identification scores ( $R^2_{\text{change}} = 4.58\%$ ,  $p = .008$ ). This relationship then mediated the relationships between parental teaching and Grades 1 and 2 letter-word identification, single-word reading and spelling rates, and phonological awareness (rhyme detection and phonological deletion). Parent-child reading was independently related to Grade 1 vocabulary ( $R^2_{\text{change}} = 5.6\%$ ,  $p = .005$ ). Thus, both home practices are relevant but to different aspects of literacy and language development.

**Keywords:** early home literacy environment, print exposure, reading development, phonological processing, parent-child reading

The secret of it all lies in the parents reading to and with the child.  
—Edmund Burke Huey, *The Psychology and Pedagogy of Reading*

Parents are encouraged to read to their children from an early age to prepare them for literacy acquisition after school entry. The U.S. Commission on Reading argued that reading to children is “the single most important activity in building the knowledge required for eventual success in reading” (Anderson, Hiebert, Scott, & Wilkinson, 1985, p. 23). Thus, the mechanism by which parent-child reading aids the child's reading acquisition is of interest. Reading acquisition is known to be fostered by several preliteracy skills that emerge in the preschool years (e.g., Adams, 1995). These are oral language skills (phonological awareness and vocabulary) and written language skills (especially letter knowledge).

Parent-child reading fosters these preliteracy skills, which provides a mechanism to explain the relationship between parent-child reading and children's own reading. In their meta-analysis, Bus, van IJzendoorn, and Pellegrini (1995) reported moderate effect sizes for the relationships between the frequency of parent-child reading and language skills ( $d = 0.67$ ), emergent literacy (letter knowledge and phonological processing;  $d = 0.58$ ), and reading achievement ( $d = 0.55$ ). Overall, parent-child reading explained around 8% of the variance in early reading skill (Scar-

borough, Dobrich, & Hager, 1991). The effects were larger in younger samples, suggesting that the relationship is stronger around the emergence of reading. Bus et al. also found that the effects of parent-child reading did not differ by socioeconomic status (SES). Even in families with low SES and low literacy level and with few other incentives to become literate, engagement in shared reading had a positive impact on children's language and literacy outcomes.

Sénéchal, Le Fevre, and colleagues examined the association between more formal literacy teaching practices (e.g., direct teaching of letter names and sounds) and these preliteracy skills and subsequent reading (Sénéchal & Le Fevre, 2002; Sénéchal, Le Fevre, Thomas, & Daley, 1998). Sénéchal et al. (1998) found that informal reading and formal teaching were independent factors in the early home environment. Further, formal teaching was more relevant in children's reading acquisition than informal parent-child reading (Sénéchal & Le Fevre, 2002; Sénéchal et al., 1998), yet far fewer parents engage in teaching than in storybook reading with preschool children (Wood, 2002).

Sénéchal and Le Fevre (2002) proposed a causal model of the relationships between preschool parent-child reading and teaching practices, the preliteracy factors related to early reading, and subsequent reading development over the first 3 years of school. The model was based on a theoretical view of learning to read that emphasized the facilitating roles of phonological awareness and letter knowledge in word identification, and of language skills in comprehension processes in reading (e.g., Adams, 1995). Sénéchal and Le Fevre proposed a direct path from frequency of parent-child book reading to receptive language (receptive vocabulary and listening comprehension) skills at the start of first grade, which in turn predicted reading (word reading and passage comprehen-

---

*Editor's Note.* Linda Baker served as action editor for this article.—KRH

Michelle Hood, Elizabeth Conlon, and Glenda Andrews, School of Psychology, Griffith University, Queensland, Australia.

Correspondence concerning this article should be addressed to Michelle Hood, School of Psychology, Griffith University, Gold Coast, PMB 50 Gold Coast Mail Centre 9726, Australia. E-mail: michelle.hood@griffith.edu.au

sion) at the end of third grade. In contrast, there was a direct path from the frequency of parental literacy teaching to emergent literacy skills (consonant–vowel–consonant word decoding, invented spelling, and letter name knowledge) at the start of first grade, which, in turn, was related to reading at the end of first grade. Sénéchal and Le Fevre found no evidence for direct paths from either parental practice to phonological awareness at the start of first grade. Thus, there were distinct pathways from each preschool home literacy practice to different underlying causal components of reading. Parent teaching was relevant in fostering early word identification, whereas parent–child reading was initially relevant in fostering language development. However, by the third grade, once reading involved both word decoding and reading comprehension, these distinct language and emergent literacy paths were both relevant in mature reading.

In the current study, we tested and extended Sénéchal and Le Fevre's (2002) model of the direct and mediated paths between early home literacy practices, children's emergent literacy skills, and subsequent literacy and language skills in a 3-year longitudinal study (preschool to Grade 2) using a different cultural and SES sample (an Australian low- to middle-class sample). We focused on predictors of single-word identification (Woodcock's, 1997, letter–word identification accuracy and a 1-min timed measure of single-word identification rate) rather than comprehension because of the younger age of our sample at the final phase. We also extended the model by examining predictors of spelling development. Both Home Literacy Environment Questionnaire items plus a children's Title Recognition Test (TRT) were used to assess the early home literacy environment. This was intended to capture a wider range of potentially important aspects of the home environment.

Bus et al. (1995) reported that relationships between preliterate and reading outcome measures did not differ depending on whether the home literacy environment was measured using a single question (e.g., the frequency of book reading) or a composite of questionnaire items (e.g., the frequency of book reading, the number of children's books owned, and the frequency of library visits). Sénéchal et al. (1998) argued against the use of questionnaire items on the basis of difficulty estimating the frequency of literacy activities, strong social desirability, and their failure to meet adequate psychometric criteria. The TRT and the Author Recognition Test (ART) are considered more objective measures of parent–child reading that overcome some of these methodological flaws. However, Sénéchal, Le Fevre, Hudson, and Lawson (1996) found that questionnaire items and TRT and ART scores were significantly related, suggesting that they both assess the construct of home literacy environment. TRTs and ARTs consist of real book titles (TRT) and author's names (ART), as well as foils (made-up titles or names). The scoring is based on signal detection theory. The number of foils "recognized" measures response bias, which is used to correct the number of real items recognized. Olson, Wise, Johnson, and Ring (1997) argued that ART and TRT have high error variance because of guessing and the ability to recognize a title or author that has not actually been read. However, Echols, West, Stanovich, and Zehr (1996) argued that many activities that result in recognizing titles or authors that have not been read are themselves indicative of a literate environment (e.g., seeing them in bookstores, libraries, or newspapers).

### Parent–Child Reading and Children's Literacy Development

Greater parent–child reading is consistently associated with more advanced language skills. Between 6.4% and 13% of the variance in language (generally measured as receptive vocabulary) was explained by parent–child reading, measured either as the age that reading first began (Burgess, Hecht, & Lonigan, 2002), a TRT/ART composite (Sénéchal & Le Fevre, 2002; Sénéchal et al., 1998), or a composite of Home Literacy Environment Questionnaire items and TRT (Fritjers, Barron, & Brunello, 2000), with the latter composite accounting for the most variance. These relationships were independent of age, earlier oral skill, phonological awareness, letter knowledge, parental education and personal reading habits, and literacy teaching practices. Meyer, Wardrop, Stahl, and Linn (1994) reported a similar positive relationship between the frequency of kindergarten teachers' reading to children and the children's language. Only Evans, Shaw, and Bell (2000) failed to find a significant relationship.

In contrast, most studies failed to find significant relationships between parent–child reading and children's letter knowledge (Evans et al., 2000; Sénéchal & Le Fevre, 2002; Sénéchal et al., 1998) or phonological awareness (Baker, Fernandez-Fein, Scher, & Williams, 1998; Cunningham & Stanovich, 1993; Evans et al., 2000; Foy & Mann, 2003; Sénéchal & Le Fevre, 2002; Sénéchal et al., 1998) that were independent of age, IQ, oral language, and parental print exposure and education. Significant relationships were found in a couple of studies (Burgess et al., 2002; Fritjers et al., 2000); however, Fritjers et al. (2000) found that phonological awareness perfectly mediated the relationship between parent–child reading and children's letter knowledge.

Studies examining the relationship of parent–child reading and the child's own subsequent reading produced similarly mixed results. The amount of parent–child reading did not differentiate precocious early readers from nonreaders (Stainthorp & Hughes, 2000) or good first grade readers from poor readers (Elbro, Borström, & Petersen, 1998). Estimates of the variance explained in children's Grade 1 word reading by parent–child reading vary from nonsignificant (Sénéchal & Le Fevre, 2002) through 3.2% (Burgess et al., 2002) to 34.9% (Cunningham & Stanovich, 1993). One explanation as to why Cunningham and Stanovich (1993) found that parent–child reading (TRT scores) explained such a large percentage of the variance in Grade 1 reading is that they had a small sample ( $N = 26$ ) that might not have been representative of the general beginner reader population.

Although Sénéchal and Le Fevre (2002) did not find that parent–child reading (TRT/ART scores) predicted Grade 1 reading, they did find that it predicted Grade 3 reading. However, this relationship was perfectly explained by shared variance with Grade 1 receptive language. Thus, the effect of earlier parent–child reading on later reading was indirect.

Cunningham and Stanovich (1993) also examined the relationship between parent–child reading and spelling. They found that TRT scores explained 20%–40% of the variance in spelling, after partialling out variance due to phonological processing. However, Evans et al. (2000) found that both phonological processing and earlier letter knowledge accounted for significant variance in Grade 1 and Grade 2 spelling. Thus, it is possible that earlier letter knowledge mediates any relationship between parent–child read-



ing and children's spelling that is not accounted for by phonological processing.

### Parent Literacy Teaching and Children's Literacy Development

Although limited, the research examining the role of more formal parent teaching practices consistently shows that engaging preschool children in more formal letter-based activities is predictive of children's own emerging literacy skills. Children who were precocious early readers or who had significantly better letter knowledge and emergent word identification skills than their peers had parents who taught them letters and writing skills (Durkin, 1966; Haney & Hill, 2004; Jackson, Donaldson, & Cleland, 1988). Parental teaching accounted for up to 10% of the variance in children's letter knowledge, after controlling for age, cognitive ability, phonological awareness, parent education, and storybook reading (Evans et al., 2000; Sénéchal & Le Fevre, 2002). According to Sénéchal and Le Fevre's (2002) model, the relationship between preschool parent teaching and children's later reading is mediated by this earlier relationship with emergent literacy skills.

Previous studies found mixed results regarding the relationship between parent teaching and language skills. Parent teaching was not directly related to phonological awareness, after controlling for letter knowledge and vocabulary (Foy & Mann, 2003; Sénéchal & Le Fevre, 2002). Haney and Hill (2004) found a trend toward children whose parents taught literacy skills having more advanced vocabulary (composite of receptive and expressive) than those whose parents did not ( $p = .07$ ). When they examined specific teaching practices, they found that teaching letter sounds was related to significantly more advanced vocabulary. However, Sénéchal and Le Fevre (2002) found that teaching did not account for unique variance in receptive language, after variance due to grade level, phonological awareness, emergent literacy, and parent education was accounted for. Evans et al. (2000) also failed to find a significant relationship.

### The Current Study

Sénéchal and Le Fevre's (2002) model formed the theoretical basis for predictions tested in the current study. Studies that examined spelling development (Cunningham & Stanovich, 1993; Evans et al., 2000) were used to extend the model to include predictions about spelling development. The effects of age, gender, memory, and non-verbal ability were controlled. It was expected that parent-child reading (measured as a composite of Home Literacy Environment Questionnaire items and TRT) would be directly related to receptive vocabulary and that this relationship would mediate any relationship between parent-child reading and children's subsequent word reading accuracy and rate. The frequency of parental teaching was expected to be directly related to preschool letter knowledge, and this relationship was expected to mediate any relationship between parental teaching and subsequent development in word reading accuracy and rate. Parent-child reading was expected to be related to spelling rate. We examined the extent to which letter knowledge mediated this relationship. The relationship between parent literacy teaching and spelling development was examined, but no specific predictions were made because of the lack of prior evidence. Consistent with Sénéchal and Le Fevre's model, neither parental reading nor teaching was

expected to be directly related to preschool phonological awareness, independently of letter knowledge and vocabulary. Any relationship between parent-child reading and letter knowledge was expected to be explained by overlapping variance with control measures, such as age, IQ, and vocabulary. Overlapping variance with the control measures and letter knowledge was expected to explain any relationship between parental teaching and vocabulary.

### Method

#### *Participants*

The initial sample comprised 143 preschool children (79 male participants and 64 female participants; mean age = 5.36 years,  $SD = 0.29$ ) who met the selection criteria of no serious developmental or intellectual impairments (parent and preschool teacher report) and English as their main language at home. The language requirement resulted in a largely Caucasian sample, with a few children of Asian or indigenous ethnicity. The three preschools from which the children were drawn were mainly composed of low- to middle-class families. Preschool is a noncompulsory year prior to school entry; however, during the period of this research, 92.6% of children who attended the school in Grade 1 attended preschool there (Education Queensland, 2006). Children attended preschool for 12.5 hr spread over 2.5 days each week. There is no formal instruction in reading or writing; however, children are encouraged to write their names on art works, are regularly read to, and engage in games to promote phonological awareness, such as rhyming and clapping out syllables. Teachers will also assist children who want to write words (e.g., write the word for the child to copy). Formal instruction in reading and writing begins in Grade 1.

In Term 2 of Grade 1 (6–8 months later), 123 of those children (68 male participants, 55 female participants; mean age = 5.95 years,  $SD = 0.30$ ) were available for retesting, and 12 months later in Term 2 of Grade 2, 105 remained available for the third phase of testing (60 male participants, 45 female participants; mean age = 7.02 years,  $SD = 0.29$ ). Thus, the attrition rate from preschool to Grade 2 was 26.6%, which is similar to the attrition rates reported in other longitudinal studies over similar periods (Leseman & de Jong, 1998; Sénéchal & Le Fevre, 2002). There were no significant differences between children who completed the study and those who did not on the early home literacy measures.

#### *Procedure*

School and written informed parental consent were obtained. All testing was conducted individually in a quiet room at the child's school over 6–7 sessions of 5–15 min each. Sessions were on different days over several weeks so fatigue was not a problem. Standardized testing procedures were followed for all commercially available tests. Administration procedures for other tests are described in the *Materials* section.

At the start of the fourth term of preschool, parents completed the Home Literacy Environment Questionnaire, which included a TRT. The children completed the letter-word identification, phonological processing, and memory tasks administered by Michelle Hood. In the Grade 1 and Grade 2 phases, these child measures were repeated, and children also completed the reading and spelling rate measures. Reading and spelling rates were always con-

ducted in separate sessions as they used the same word lists. In the Grade 1 phase, nonverbal ability and receptive vocabulary were also assessed.

### Materials

#### Early Home Literacy Environment

**TRT.** The parent-completed children's TRT included 20 popular age-appropriate children's book titles and 10 foils (see Appendix A). The titles were derived from Angus and Robertson Bookworld's (1999) 100 all-time favorite children's books and previously used TRTs (Cunningham & Stanovich, 1993; Sénéchal et al., 1998). These titles also represented current best sellers. The score was the proportion of real titles checked minus the proportion of foils checked, with negative scores entered as zero. Cronbach's alpha was .74 for the real titles, .64 for the foils, and .76 for the total scale.

**Home Literacy Environment Questionnaire.** Parents responded regarding the frequency of reading to the child, the number of children's books, the frequency of parental teaching of literacy skills, and the frequency of library visits, as well as other nonliteracy activities included as filler items (see Appendix B). Questions were based on those previously used by Sénéchal et al. (1998) and Foy and Mann (2003). The Home Literacy Environment Questionnaire also asked parents about the child's interest in being read to. This was included because a lack of interest by the child may affect the extent to which parents engage in literacy activities with the child. Child interest was significantly correlated with the frequency of parent's reading (Crain-Thoreson & Dale, 1992; Dunning, Mason, & Stewart, 1994; Olofsson & Niedersoe, 1999) and with children's letter knowledge (Fritjers et al., 2000), word and sentence reading (Olofsson & Niedersoe, 1999; Scarborough et al., 1991), and language skills (Payne, Whitehurst, & Angell, 1994; Sénéchal et al., 1998). The questionnaire also requested demographic details (e.g., age, gender, and medical and developmental history).

#### Control Measures

**Memory.** Several authors argued that memory needs to be controlled in predictive studies of reading because of its significant relationship with early reading development (Mann & Liberman, 1984; Molfese, Molfese, & Modgline, 2001; Wagner & Torgesen, 1987). The auditory-verbal short-term memory measure was the Digit Span Forward subtest from the Dyslexia Early Screening Test (Nicolson & Fawcett, 1996), which is suitable for children 4.5–6.42 years of age. Nicolson and Fawcett (1996) reported a 1-week test-retest reliability of .63 for children 5.5–6.5 years of age. We found a 6-month test-retest reliability (preschool to Grade 1) of .60. There were two trials at each digit span from two to nine digits. Testing ceased when two trials at any given span were incorrect. The memory score was the number of correct trials (maximum = 16).

**Nonverbal ability.** This was included as a measure of children's analytic intelligence and was measured with Raven's Colored Progressive Matrices (Raven, Court, & Raven, 1986). The reported Cronbach's alpha for an Australian sample (mean age = 5.5 years) was .80 (Raven et al., 1986). Our obtained alpha was .75.

### Phonological Awareness

There were two measures of phonological awareness. For each measure, *z* scores were formed and summed to give a composite phonological awareness score.

**Rhyme and alliteration detection.** The Rhymes subtest of the Cognitive Profiling System (CoPS; Singleton, Thomas, & Leedale, 1997) was used. CoPS is a computerized early screening test for children age 4.5 years and over. Fawcett, Singleton, and Peer (1998) reported prediction rates for reading risk of over 90% using the CoPS, and acceptable false-negative (12.0%–16.7%) and false-positive (around 2 %) rates. They found a strong correlation ( $r = .52$ ,  $p < .001$ ) between Rhymes performance at 5 years and single-word recognition at 8 years. We found a 6-month test-retest reliability (preschool to Grade 1) of .72 and Cronbach's alphas at each phase of .87–.93.

Participants chose a stimulus word from four choices that sounded like the target word on rhyme trials (eight trials) or that started with the same sound on alliteration trials (eight trials). On each trial, the spoken target and stimulus words were accompanied by pictures of the words on the computer screen, which remained in place until the child responded. This reduced memory demands. At preschool, the children only completed the rhyme trials. At Grades 1 and 2, they completed the alliteration trials if they correctly completed the rhyme trials. Maximum possible scores were 8 at preschool and 16 at Grades 1 and 2.

**Phonological segmentation.** The Phonemic Segmentation subtest of the Dyslexia Early Screening Test (Fawcett & Nicolson, 1996) involves deleting a phonological segment from a word read out by the experimenter (e.g., say *brain* without the /b/ to get *rain*). The deleted segment ranges from a syllable to a single phoneme within a blend, and varies across initial, medial, and final positions within the word. There were 12 trials. Fawcett and Nicolson (1996) reported a 1-week test-retest reliability of .88 for children 6.5–12 years of age. We found a 6-month (preschool to Grade 1) test-retest reliability of .62 and Cronbach's alphas of .57–.61.

### Receptive Vocabulary

The Peabody Picture Vocabulary Test—Revised Form M (Dunn & Dunn, 1981) is a graded test suitable for people 2.5–40 years of age. Participants chose one of four pictures that best illustrated the meaning of a word spoken by the examiner. There were three training items first. Dunn and Dunn (1981) reported split half reliabilities of around .80 for children 5–6.92 years of age. We found a Spearman-Brown-corrected split half reliability of .81.

### Reading

**Letter-word identification.** The Letter-Word Identification subtest of the Woodcock Diagnostic Reading Battery (Woodcock, 1997) is a graded list, beginning with selected letters (upper and lower case) and continuing with words of increasing difficulty. Reported internal consistency was .94 for participants 5–18 years of age (Woodcock, 1997). We found Cronbach's alphas ranging from .83 to .93 across the three phases.

**Reading rate.** This measured word identification fluency. Participants read as many words as possible in 1 min from a list based on the most frequent English words from Dolch's (1936) and



Kucera-Francis’s (1967) sight-word lists (see Appendix C). These words were also contained on the participating school’s sight-word lists for Grades 1–3 as well as the words that the children were expected to be learning. Words were presented in order from highest frequency/easiest in three columns on A–4 pages in size 20 Berlin Sans FB font. The dependent measure was the number correctly identified in 1 min (maximum = 120). Test–retest reliability over the 12 months from Grade 1 to Grade 2 was .79.

Spelling Rate

Children wrote to dictation as many words as possible in 2 min from the list of high frequency words used in the reading rate task (see Appendix C). The experimenter read the word, repeated it in a phrase or sentence, and if the child had not yet finished writing the word, repeated the word again. Immediately when the child finished one word, the next word was read out loud. The dependent measure was the number of correctly spelled words (maximum = 120). Test–retest reliability over the 12 months from Grade 1 to Grade 2 was .71.

Results

Literacy Practices in the Early Home Environment

Appendix B includes the percentages of parents choosing each response option for all items on the Home Literacy Environment Questionnaire. All parents reported reading to their child at least once per week, with 58.4% reading once or more per day. All families owned at least one book per child, with around 75% reporting they owned 50 or fewer. Only 7% of parents reported that their child was not interested or only slightly interested in being read to. A total of 21% reported that they never took the child to the library (although all children visited the school library weekly). Most parents reported that they taught the child the alphabet (81.2%), to write their name (76.3%), and to read (57.4 %) often or very often. Thirteen parents (9.1%) did not report frequency for teaching their child to write their name because they reported that the child could already do that. Six of those parents also did not report the frequency of teaching the alphabet for that same reason. These responses were treated as missing data, and scores were replaced with the median to retain these participants in the sample. These children did not differ significantly on the outcome measures from the children whose parents had reported a frequency of doing these activities at the median for the sample, and they performed significantly better on the outcome measures than those children whose parents reported a frequency for these activities at lower than the median. This supported our replacement of their missing data with the median. The mean score on the TRT was .29 (*SD* = .16; *M*<sub>real</sub> = .35, *M*<sub>foils</sub> = .07). In other words, after correcting for response bias, an average 29.19% of the real book titles were recognized (see Appendix A).

A principal components analysis (varimax rotation) of the Home Literacy Environment Questionnaire items and the TRT produced two factors, labeled Parental Teaching and Parental Reading (see Table 1). Frequency of library visits was excluded as it did not load clearly on either factor. The two factors accounted for 57.82% of the variance in the six remaining items. To form a composite Parental Reading measure, we first standardized reading item

Table 1  
*Principal Components Analysis of Items Measuring Early Home Reading Environment (N = 143)*

Item	Factor loadings	
	Parental teaching	Parental reading
Frequency teach alphabet	.83	.01
Frequency teach write name	.76	.12
Frequency teach reading	.72	.06
No. children’s book per child	.14	.78
Children’s TRT score	–.19	.73
Frequency reading per week	.29	.63
Eigenvalues	2.08	1.39
Percentage of variance	34.64	23.18
Kaiser–Meyer–Olkin measure	0.62	
Bartlett’s test of sphericity	$\chi^2(15) = 126.53, p < .0001$	

*Note.* The minimal factor loading required to be significant at an alpha level of .05 was .50 (Hair, Anderson, Tatham, & Black, 1998). TRT = Title Recognition Test.

scores and then summed them. As the teaching items were all measured on the same response scale, initial standardizing was not required. They were simply summed to form the composite Parental Teaching measure. These two factors were only weakly correlated (3.35% shared variance). Child Reading Interest was significantly correlated with each—more strongly so with Parental Reading than with Parental Teaching (see Table 2). However, it was not significantly correlated with any outcome measures, so it was not included in any further analyses.

Table 2 presents the zero-order correlations, and Table 3 presents the descriptive statistics for the control measures (age, gender, memory, and nonverbal ability), the literacy and language measures, and the home literacy factors. Grade 1 reading and spelling rates were square-root transformed to normalize the positively skewed distributions. These transformed variables were used in the analyses. At preschool, letter–word identification was mainly a measure of letter identification (a score of 13 corresponds to knowing only letters; scores of 14 plus require word identification).

Mediation Analyses

Baron and Kenny’s (1986) approach to testing mediation (see also recommendations of Kenny, Kashy, & Bolger, 1998; Shrout & Bolger, 2002) was used to test the extent to which indirect paths via the potential mediators (phonological awareness, letter identification, and vocabulary) explained the relationship between Parental Reading and Parental Teaching and subsequent literacy and language skills.<sup>1</sup> Figure 1 represents the general mediation model tested. Path c is the direct pathway between the home environment predictors (Parental Reading or Parental Teaching) and the language and literacy criterion measures (vocabulary, reading, and spelling). Paths a and b together represent the mediated or indirect

<sup>1</sup> Vocabulary functioned as an initial outcome measure and then as a mediator between home literacy practices and later outcome measures.

Table 2  
Zero-Order Correlations Between Early Home Literacy Practices, Control Measures, Reading, Spelling, Phonological Processing, and Vocabulary

Variable	1	2	3 <sup>a</sup>	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. TEACH	—																			
2. READ	.18*	—																		
3. Gender <sup>a</sup>	-.13	-.09	—																	
4. Age	.12	.13	-.07	—																
5. pMEM	-.13	.07	.11	.08	—															
6. lMEM	-.12	.04	.05	-.07	.60*	—														
7. 2MEM	-.05	.23*	.17	-.05	.56*	.59*	—													
8. NONVERB	.04	.20*	.14	.25**	.17	.10	.24*	—												
9. pLettWord	.24*	.17*	-.02	.28**	.17*	.23*	.16	.33**	—											
10. lLettWord	.22*	.04	-.07	.27**	.15	.24**	.18	.17	.69**	—										
11. 2LettWord	.20*	-.03	.04	.22*	.20*	.26*	.16	.21*	.56**	.73**	—									
12. lReadRate	.28*	.08	-.10	.25**	.10	.21*	.10	.17	.67**	.83**	.83**	—								
13. 2ReadRate	.19*	-.02	-.06	.18	.16	.22*	.17	.04	.48**	.81**	.81**	.80**	—							
14. lSpellRate	.25**	.12	-.17*	.26**	.09	.22**	.04	.09	.58**	.72**	.67**	.81**	.65**	—						
15. 2SpellRate	.23**	-.03	-.11	.22*	.21*	.28**	.15	.08	.47**	.70**	.71**	.75**	.78**	.71*	—					
16. pPHON	.13	.14	.05	.27**	.32**	.28**	.22*	.22*	.53**	.41**	.44**	.43**	.39**	.47**	.32**	—				
17. lPHON	.17	.11	.09	.18*	.28**	.34**	.27**	.20*	.54**	.55**	.52**	.50**	.47**	.53**	.45**	.75**	—			
18. 2PHON	.22**	.02	.04	.14	.17	.30**	.18	.19	.51**	.60**	.71**	.65**	.63**	.57**	.59**	.52**	.62**	—		
19. VOCAB	.24**	.30**	-.05	.24**	.21*	.14	.20*	.33**	.40**	.43**	.32**	.34**	.30**	.39**	.29**	.35**	.32**	.27**	—	
20. INTEREST	.17*	.32**	-.13	-.01	.04	.02	.01	.02	.06	-.03	.00	-.09	-.02	-.07	-.02	.04	-.01	.11	.13	—

Note. TEACH = parental teaching; READ = parental reading; MEM = memory; p = preschool, 1 = Grade 1; 2 = Grade 2; NONVERB = nonverbal ability; LettWord = letter-word identification; Read Rate = reading rate; Spell Rate = spelling rate; PHON = phonological awareness; VOCAB = receptive vocabulary; INTEREST = parent report of child interest in being read to.

<sup>a</sup> Categorical variable, Kendall's tau reported.

\*  $p < .05$ . \*\*  $p < .01$ .



Table 3  
*Descriptive Statistics of Control Measures, Predictors, and Outcome Literacy*

Measure	<i>M</i>	<i>SD</i>	Range
Preschool age (years)	5.36	0.29	4.83, 6.17
Grade 1 age (years)	5.95	0.30	5.41, 6.75
Grade 2 age (years)	7.02	0.29	6.50, 7.59
Preschool memory	5.19	1.44	2.00, 9.00
Grade 1 memory	5.90	1.42	2.00, 11.00
Grade 2 memory	6.74	1.55	4.00, 10.00
Nonverbal ability <sup>a</sup>	108.43	10.39	81.00, 135.00
Receptive vocabulary <sup>a</sup>	101.57	12.15	72.00, 129.00
Preschool phonological awareness <sup>a</sup>	0.20	1.76	-3.57, 3.70
Grade 1 phonological awareness <sup>a</sup>	0.15	1.64	-2.99, 3.89
Grade 2 phonological awareness <sup>a</sup>	0.00	1.62	-5.56, 2.33
Preschool letter-word identification	10.23	3.05	4.00, 20.00
Grade 1 letter-word identification	14.63	3.69	5.00, 28.00
Grade 2 letter-word identification	27.59	7.18	12.00, 46.00
Grade 1 reading rate	18.06	14.33	0.00, 78.00
Grade 2 reading rate	57.13	24.52	8.00, 120.00
Grade 1 spelling rate	8.05	5.72	0.00, 30.00
Grade 2 spelling rate	22.42	7.23	3.00, 48.00
Parental teaching	12.51	1.97	7.00, 15.00
Parental reading	15.57	5.78	5.15, 30.60

<sup>a</sup> Standardized scores.

path from the predictor to the criterion measure via the potential mediators (letter identification and vocabulary).

Baron and Kenny (1986) identified three necessary conditions for mediation. First, the predictor must account for significant variation in the hypothesized mediator (see Path a, Figure 1). Hierarchical multiple regression was used to examine this relationship between the home environment factors and the potential mediators. Note that in all of these analyses, variance explained by the control measures was partialled out first. Second, the mediator must account for significant variation in the criterion (see Path b, Figure 1). This was tested by regressing the criterion measure on both the predictor and mediator. In line with the recommendations of Kenny et al. (1998), the second condition is satisfied if the mediator explains a significant independent component of variance in the criterion. The final condition (see Path c', Figure 1) is that when the relationships between the predictor and the mediator (Path a) and between the mediator and the criterion (Path b) are accounted for, a previously significant relationship between the predictor and the criterion (Path c) becomes nonsignificant (complete mediation) or is substantially reduced (partial mediation). Standardized beta coefficients were examined to determine the extent to which the independent contribution of the predictor was reduced. Finally, we applied Sobel's (1982) test<sup>2</sup> (using Preacher & Leonardelli's, 2001, interactive calculation tool) to determine whether this mediated path (Paths a and b) significantly differed from zero.

In all of the subsequent hierarchical regression analyses, control measures were entered at Step 1, prior to home literacy environment predictors, to initially partial out the variance they explained in the specific criterion measure. Control and predictor variables that were not significantly bivariately correlated with that criterion measure (see Table 2) were not included. Results for the control measures are shown in the regression summary tables but are not

detailed further in text. Results with the preschool emergent literacy skills (phonological awareness and letter identification) as criterion measures are presented first, followed by those with receptive vocabulary and then the Grade 1 and Grade 2 outcome measures as the criterion measures.

### *Prediction of Preschool Emergent Literacy Skills by Parental Literacy Practices*

On the basis of Sénéchal and Le Fevre's (2002) model, neither Parental Reading nor Parental Teaching was expected to show a significant direct relationship with Phonological Awareness. Nonsignificant zero-order correlations (see Table 2) confirmed this. Thus, there were no direct paths between the parent practices and early phonological awareness.

Parental Teaching was expected to be directly related to Preschool Letter-Word Identification. The hierarchical regression analyses are summarized in Table 4a. At Step 2, Parental Teaching explained an additional 7.6% of the variance in Preschool Letter-Word Identification (see significant Path c, Figure 1) over that accounted for by the control measures at Step 1. Sénéchal and Le Fevre (2002) found the direct relationship between Parental Teaching and emergent literacy was not explained by Vocabulary or Phonological Awareness. We tested these potential indirect relationships using more formal mediation analysis. Phonological Awareness was not significantly correlated with Parental Teaching (see Table 2) so was not considered further as a mediator (i.e., nonsignificant Path a). However, Vocabulary was significantly correlated with Parental Teaching (see Table 2), and after partialling out variance due to the control measures, Parental Teaching independently explained 5.6% of the variance in Vocabulary,  $F(1, 117) = 8.37, p = .005$  (significant Path a). When Letter-Word Identification was regressed on both Parental Teaching and Vocabulary, Vocabulary independently explained 4.1% of the variance (significant Path b). With this mediated path via Vocabulary accounted for, Parental Teaching still independently explained a significant 4.58% of the variance in Letter-Word Identification. Sobel's test results indicated that the mediated path via Vocabulary did not quite reach significance ( $z = 1.94, p = .051$ ). Thus, consistent with Sénéchal and Le Fevre's model, there were direct paths from both Parental Teaching and Vocabulary to Preschool Letter-Word Identification.

Parental Reading was significantly correlated with Preschool Letter-Word Identification (see Table 2). However, once differences in age, memory, and nonverbal ability were accounted for, this relationship became nonsignificant. Consistent with Sénéchal and Le Fevre's (2002) model, there was no direct path from Parental Reading to Letter-Word Identification.

### *Prediction of Receptive Vocabulary by Parental Literacy Practices*

Sénéchal and Le Fevre's (2002) model predicted that Parental Reading would be related to Vocabulary, independently of rela-

<sup>2</sup> Sobel's formula:  $z \text{ value} = a \times b / \text{SQRT}(b^2 \times s_a^2 + a^2 \times s_b^2)$ , where SQRT = square root,  $a$  = raw regression coefficient ( $B$ ) for Path a,  $s_a$  = standard error of  $a$  ( $SE B$ ),  $b$  = raw regression coefficient for Path b, and  $s_b$  = standard error of  $b$ .

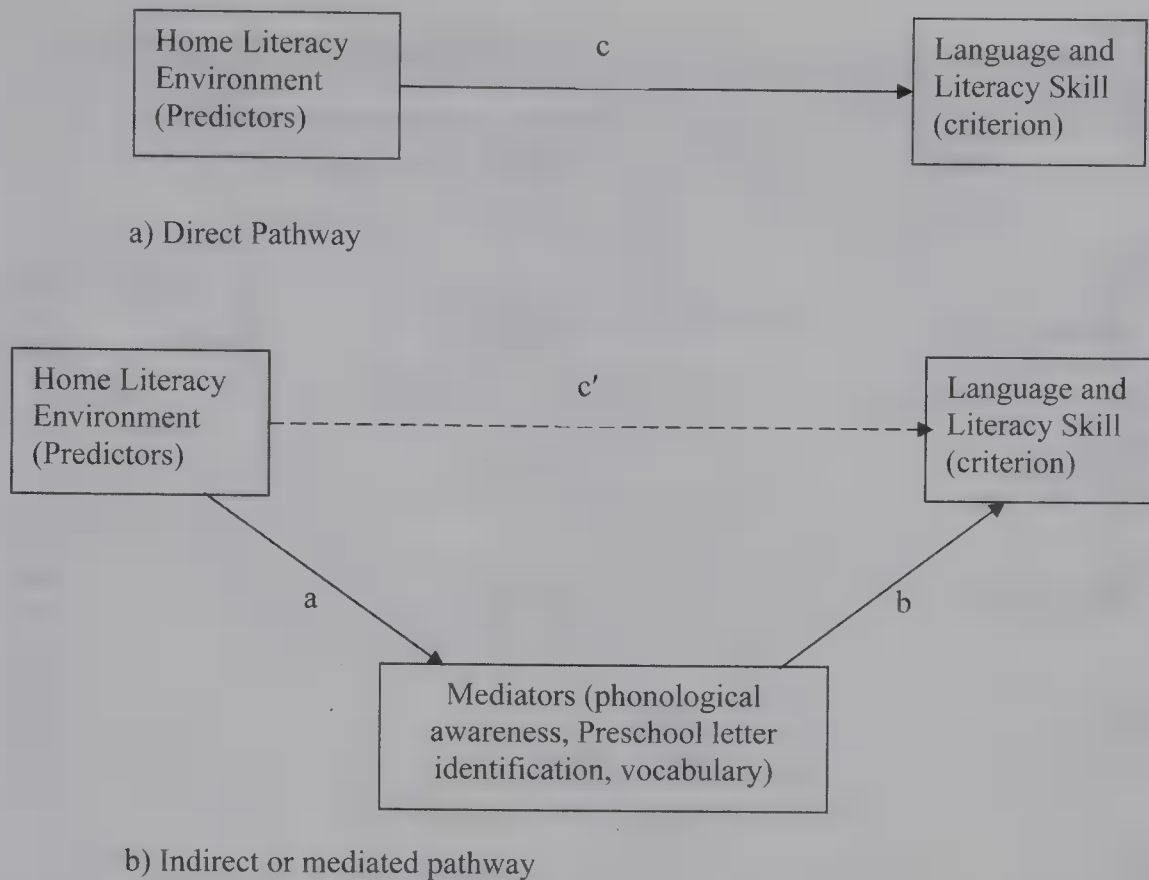


Figure 1. Generic mediation model being tested (on the basis of Baron & Kenny, 1986).

tionships with emergent literacy and phonological awareness. Table 4b summarizes these analyses. At Step 2, Parental Reading independently explained 5.6% of the variance in Vocabulary (significant Path c). Neither Preschool Letter–Word Identification nor Phonological Awareness met Baron and Kenny’s (1986) first condition for a mediator (significant Path a from predictor to mediator), after partialling out variance explained by the control measures (see Table 4a). Thus, there was evidence only for a significant direct path from Parental Reading to Vocabulary that was independent of emergent literacy skills and phonological awareness.

Parental Teaching was also significantly correlated with Vocabulary (see Table 2). With variance due to the control measures explained, Parental Teaching explained 4.5% of the variance in Vocabulary (see Table 4b). Preschool Letter–Word Identification was a potential mediator as it was significantly predicted by Parental Teaching (see Table 4a; Path a). When it was entered in the regression analysis with Parental Teaching, Preschool Letter–Word Identification independently explained 5.6% of the variance in Vocabulary, and Parental Teaching no longer explained a significant percentage of the variance. Thus, there was complete mediation by Letter–Word Identification. This mediated path was significantly different from zero ( $z = 2.13, p = .033$ ). Thus, as predicted by Sénéchal and Le Fevre’s (2002) model, there was only an indirect path from Parental Teaching via Preschool Letter–Word Identification to Vocabulary.

#### *Prediction of Mid-Grade 1 Literacy Outcomes by Parental Literacy Practices*

The criterion measures in this set of mediation analyses were Grade 1 Phonological Awareness, Letter–Word Identifi-

cation, Reading Rate, and Spelling Rate. Parental Reading showed no significant zero-order correlations with any of the Grade 1 outcome measures (see Table 2); thus, no further analyses involving it were conducted. Parental Teaching was significantly correlated with all Grade 1 outcome measures, except Phonological Awareness. Thus, no further analyses predicting Grade 1 Phonological Awareness were conducted, and no paths from parental practices to Grade 1 Phonological Awareness were supported in the model. Table 5 summarizes the separate hierarchical regression analyses conducted for the other Grade 1 criterion measures. Preschool Letter–Word Identification was a potential mediator in all analyses based on significant zero-order correlations with each criterion measure (see Table 2) and significant prediction by Parental Teaching (see Table 4a).

Table 5a summarizes the regressions on Grade 1 Letter–Word Identification. At Step 2, Parental Teaching explained 6.4% of the variance in Grade 1 Letter–Word Identification. When Preschool Letter–Word Identification was also entered at Step 2, it independently explained 37.7% of the variance. Parental Teaching no longer explained a significant percentage of the variance, consistent with complete mediation by Preschool Letter–Word Identification. This mediated path was significant ( $z = 3.32, p = .0009$ ). Thus, there was no direct path from Parental Teaching to Grade 1 Letter–Word Identification—only an indirect path via earlier Letter–Word Identification.

A similar pattern of results was obtained with Grade 1 Reading Rate as the criterion (see Table 5b). At Step 2, Parental Teaching explained a further 7.9% of the variance in Grade 1 Reading Rate. When Preschool Letter–Word was also included in the regression analysis, it independently explained 29.5% of the variance, and the



Table 4  
*Hierarchical Regression Analyses for Predicting Preschool Letter-Word Identification and Grade 1 Receptive Vocabulary*

Variable	<i>B</i> ( <i>SE B</i> )	$\beta$	<i>p</i>
(a) Criterion measure = Preschool letter-word identification ( <i>N</i> = 122)			
Step 1.			
Age	1.76 (0.87)	.18	.044*
Preschool Memory	0.25 (0.18)	.12	.161
Nonverbal	0.18 (0.06)	.27	.003*
	$R^2 = .153^*$ , $F(3, 118) = 7.08$ , $p < .0001$		
Step 2.			
Parent teach	0.43 (0.13)	.28	.001*
	$\Delta R^2 = .076^*$ , $F(1, 117) = 11.59$ , $p = .001$		
Step 3.			
Parent teach	0.34 (0.13)	.23	.008*
Vocabulary	0.06 (0.02)	.23	.011*
	$\Delta R^2 = .041^*$ , $F(1, 116) = 6.60$ , $p = .011$		
Step 2.			
Parent read	0.12 (0.12)	.09	.307
	$\Delta R^2 = .008$ , $p = .291$		
(b) Criterion measure = Vocabulary ( <i>N</i> = 122)			
Step 1.			
Age	6.49 (3.47)	.17	.064
Nonverbal	0.77 (0.23)	.29	.001*
	$R^2 = .136^*$ , $F(2, 119) = 9.40$ , $p < .0001$		
Step 2.			
Parent read	1.30 (0.46)	.24	.005*
	$\Delta R^2 = .056^*$ , $F(1, 118) = 8.12$ , $p = .005$		
Step 2.			
Parent teach	1.29 (0.51)	.21	.012*
	$\Delta R^2 = .045^*$ , $F(1, 118) = 6.46$ , $p = .012$		
Step 3.			
Parent teach	0.88 (0.51)	.14	.09
Preschool letter-word identification	1.06 (0.36)	.27	.004*
	$\Delta R^2 = .056^*$ , $F(1, 117) = 8.55$ , $p = .004$		

Note. The contributions of the control measures beyond Step 1 are not included in any of the summaries in the interests of economy of presentation and because they are not the focus of the analysis

\*  $p < .05$ .

percentage of variance explained by Parental Teaching became nonsignificant. This mediated path was significant ( $z = 3.23$ ,  $p = .001$ ). Thus, any relationship between Parental Teaching and Grade 1 Reading Rate was also completely mediated by Preschool Letter-Word Identification.

Similar results were also obtained in the regression on Grade 1 Spelling Rate (see Table 5c). At Step 2, Parental Teaching explained 5.3% of the variance in Grade 1 Spelling Rate. Preschool Letter-Word Identification independently explained a further 18.9% of the variance in Grade 1 Spelling Rate and reduced the variance explained by Parental Teaching to a nonsignificant level. This mediated path was significant ( $z = 3.00$ ,  $p = .003$ ). Thus, the only path supported was a mediated path from Parental Teaching via Preschool Letter-Word Identification to Grade 1 Spelling.

### *Prediction of Grade 2 Phonological Awareness, Reading, and Spelling*

Parental Reading showed no significant zero-order correlations with any Grade 2 outcome measures (see Table 2) and was not included in further analyses. However, there were significant correlations between Parental Teaching and all Grade 2 criterion measures. Preschool Letter-Word Identification was a potential mediator of these relationships.

Prediction of Grade 2 Phonological Awareness is detailed in Table 6a. At Step 2, Parental Teaching explained an additional 7.4% of the variance in Grade 2 Phonological Awareness. When both Parental Teaching and Preschool Letter-Word Identification were entered at Step 2, Preschool Letter Word Identification explained 15.29% of the variance, and the contribution of Parental

Table 5  
*Hierarchical Regression Analyses for Predicting Grade 1 Reading and Spelling*

Variable	<i>B</i> ( <i>SE B</i> )	$\beta$	<i>p</i>
(a) Criterion measure = Grade 1 letter-word identification ( <i>N</i> = 123)			
Step 1.			
Age	3.59 (0.98)	.31	<.0001*
Grade 1 memory	0.63 (0.20)	.26	.002*
$R^2 = .153^*$ , $F(2, 120) = 10.87$ , $p < .0001$			
Step 2.			
Parent teach	0.46 (0.15)	.26	.002*
$\Delta R^2 = .064^*$ , $F(1, 119) = 9.79$ , $p = .002$			
Step 3.			
Parent teach	0.10 (0.11)	.06	.363
Preschool letter-word identification	0.82 (0.08)	.69	<.0001*
$\Delta R^2 = .377^*$ , $F(1, 118) = 109.99$ , $p < .0001$			
(b) Criterion measure = Grade 1 reading rate ( <i>N</i> = 121)			
Step 1.			
Age	1.58 (0.50)	.27	.002*
Grade 1 memory	0.27 (0.10)	.23	.009*
$R^2 = .118^*$ , $F(2, 120) = 8.05$ , $p = .001$			
Step 2.			
Parent teach	0.25 (0.07)	.28	.001*
$\Delta R^2 = .079^*$ , $F(1, 119) = 11.70$ , $p = .001$			
Step 3.			
Parent teach	0.10 (0.06)	.11	.126
Preschool letter-word identification	0.36 (0.04)	.61	<.0001*
$\Delta R^2 = .295^*$ , $F(1, 118) = 68.61$ , $p < .0001$			
(c) Criterion measure = Grade 1 spelling rate ( <i>N</i> = 123)			
Step 1.			
Age	1.02 (0.31)	.28	.001*
Grade 1 memory	0.18 (0.06)	.24	.004*
Gender <sup>a</sup>	-0.46 (0.18)	-.21	.012*
$R^2 = .174^*$ , $F(3, 119) = 8.35$ , $p < .0001$			
Step 2.			
Parent teach	0.13 (0.05)	.23	.005*
$\Delta R^2 = .053^*$ , $F(1, 118) = 8.03$ , $p = .005$			
Step 3.			
Parent Teach	0.05 (0.04)	.09	.217
Preschool letter-word identification	0.18 (0.03)	.49	<.0001*
$\Delta R^2 = .189^*$ , $F(1, 118) = 11.81$ , $p = .001$			

<sup>a</sup> Male = 1, Female = 0.

\*  $p < .05$ .

Teaching became nonsignificant. This mediated path was significant ( $z = 2.90$ ,  $p = .004$ ). Grade 1 Phonological Awareness was then added at Step 3. With it added, the mediated path remained significant, explaining 2.69% of the variance in Grade 2 Phonological Awareness. Grade 1 Phonological Awareness independently explained a further 11.7% of the variance. Thus, an indirect path from Parental Teaching via Preschool Letter-Word Identification to Grade 2 Phonological Awareness, which was independent of earlier phonological awareness, was supported in the model.

With Grade 2 Letter-Word Identification as the criterion measure, Parental Teaching explained an additional 4.63% of

the variance at Step 2 (see Table 6b). With both Preschool Letter-Word Identification and Parental Teaching entered at Step 2, Letter-Word Identification independently explained 17.06% of the variance, but the independent contribution of Parental Teaching became nonsignificant. When Grade 1 Letter-Word Identification was added at Step 3, it independently explained 37.5% of the variance, and Preschool Letter-Word Identification no longer explained a significant independent percentage of the variance. Thus, any variance in Grade 2 Letter-Word Identification explained by Parental Teaching overlapped with that already accounted for by the mediated path from Preschool to Grade 1 Letter-Word Identification. There



Table 6  
*Hierarchical Regression Analyses Predicting Grade 2 Phonological Awareness, Reading, and Spelling*

opening

Variable	B (SE B)	β	p
(a) Criterion measure = Grade 2 phonological awareness (N = 105)			
Step 1.			
Grade 1 memory	0.34 (0.11)	.30	.002*
	R <sup>2</sup> = .09*, F(1, 101) = 10.02, p = .002		
Step 2.			
Parent teach	0.23 (0.08)	.28	.004*
	ΔR <sup>2</sup> = .074*, F(1, 100) = 8.84, p = .004		
Step 2.			
Parent teach	0.13 (0.07)	.15	.096
Preschool letter–word identification	0.24 (0.05)	.42	<.0001*
	ΔR <sup>2</sup> = .227*, F(2, 99) = 16.47, p < .0001		
Step 3.			
Parent teach	0.09 (0.07)	.10	.206
Preschool letter–word identification	0.12 (0.53)	.20	.034*
Grade 1 phonological awareness	0.43 (0.10)	.44	<.0001*
	ΔR <sup>2</sup> = .117*, F(1, 98) = 20.30, p < .0001		
(b) Criterion measure = Grade 2 letter–word identification (N = 106)			
Step 1.			
Age	4.60 (2.33)	.19	.051
Grade 1 memory	1.16 (0.47)	.23	.015*
Nonverbal	0.21 (0.15)	.13	.182
	R <sup>2</sup> = .128*, F(3, 102) = 4.98, p = .003		
Step 2.			
Parent teach	0.82 (0.35)	.22	.02*
	ΔR <sup>2</sup> = .046*, F(1, 101) = 5.63, p < .0001		
Step 2.			
Parent teach	0.37 (0.32)	.10	.255
Preschool letter–word identification	1.21 (0.24)	.48	<.0001*
	ΔR <sup>2</sup> = .217*, F(2, 100) = 16.53, p < .0001		
Step 3.			
Parent Teach	0.28 (0.27)	.07	.305
Preschool letter–word identification	−0.02 (0.27)	−.01	.942
Grade 1 letter–word identification	1.48 (0.22)	.69	<.0001*
	ΔR <sup>2</sup> = .204*, F(1, 99) = 44.82, p < .0001		
(c) Criterion measure = Grade 2 reading rate (N = 104)			
Step 1.			
Grade 1 memory	3.94 (1.63)	.24	.017*
	ΔR <sup>2</sup> = .056*, F(1, 100) = 5.88, p = .017		
Step 2.			
Parent teach	3.80 (1.21)	.30	.002*
	ΔR <sup>2</sup> = .086*, F(1, 99) = 9.93, p = .002		
Step 2.			
Parent teach	2.07 (1.12)	.16	.068
Preschool letter–word identification	3.95 (0.76)	.46	<.0001*
	ΔR <sup>2</sup> = .273*, F(2, 98) = 19.96, p < .0001		
Step 3.			
Parent Teach	0.51 (0.73)	.04	.491
Preschool letter–word identification	−0.37 (0.61)	−.04	.54
Grade 1 read rate	12.15 (1.03)	.85	<.0001*
	ΔR <sup>2</sup> = .397*, F(1, 97) = 140.31, p < .0001		

Table 6 (continued)

Variable	B (SE B)	$\beta$	p
(d) Criterion measure = Grade 2 spelling rate (N = 105)			
Step 1.			
Age	5.27 (2.29)	.21	.024*
Grade 1 memory	1.38 (0.47)	.27	.004*
$R^2 = .123^*$ , $F(2, 102) = 7.13$ , $p = .001$			
Step 2.			
Parent teach	0.97 (0.34)	.26	.006*
$\Delta R^2 = .064^*$ , $F(1, 101) = 7.95$ , $p = .006$			
Step 2.			
Parent teach	0.61 (0.34)	.16	.072
Preschool letter-word identification	0.91 (0.24)	.36	<.0001*
$\Delta R^2 = .169^*$ , $F(2, 100) = 11.94$ , $p < .0001$			
Step 3.			
Parent teach	0.29 (0.28)	.08	.303
Preschool letter-word identification	0.20 (0.22)	.08	.376
Grade 1 spell rate	4.20 (0.60)	.62	<.0001*
$\Delta R^2 = .237^*$ , $F(1, 99) = 49.79$ , $p < .0001$			

\*  $p < .05$ .

was a further direct path from Grade 1 to Grade 2 Letter-Word Identification.

Results were similar for the other criterion measures. At Step 2, Parental Teaching explained an additional 8.58% of the variance in Grade 2 Reading Rate (see Table 6c). Adding Preschool Letter-Word Identification at Step 2 independently explained a further 18.75% of the variance, and the independent contribution of Parental Teaching became nonsignificant. At Step 3, Grade 1 Reading Rate independently explained a further 39.69% to the variance and reduced the contribution of Preschool Letter-Word Identification to nonsignificance. Thus, a mediated path from Parental Teaching via Preschool Letter-Word Identification to Grade 1 Reading Rate and from there a direct path to Grade 2 Reading Rate was supported.

At Step 2, Parental Teaching independently explained 6.4% of the variance in Grade 2 Spelling Rate (see Table 6d). Preschool Letter-Word Identification completely mediated this relationship and independently explained a further 10.5% of the variance. However, when Grade 1 Spelling Rate was also added, the contribution of Preschool Letter-Word Identification became nonsignificant. Grade 1 Spelling Rate independently explained 23.72% of the variance in Grade 2 Spelling Rate. Thus, there was only a mediated path from Parental Teaching via Preschool Letter-Word Identification to Grade 1 Spelling Rate, which was then directly related to Grade 2 Spelling Rate.

### Discussion

We examined the extent to which preschool home literacy teaching and reading practices account for significant variance in emergent literacy skills (phonological awareness and letter identification) at preschool and in language and literacy skills in Grade 1 and Grade 2. On the basis of results of previous studies and in particular the model proposed by Sénéchal and Le Fevre (2002), we examined the direct and mediated relationships between these

home literacy factors and the outcome measures, when variance shared with age, gender, nonverbal ability, and memory was controlled. The results indicate that parent-child reading and parent literacy teaching practices were only weakly related factors in the early home environment and that they showed different relationships to language and literacy outcomes. The model in Figure 2 summarizes our overall findings. This not only confirms the generalizability of Sénéchal and Le Fevre's model to an Australian sample drawn from schools servicing low- to middle-SES families but extends it to include both accuracy and rate of word reading as well as spelling rate.

The home environments of the children prior to school entry included a number of literacy activities. In general, the findings were consistent with previous studies that used samples from a variety of SES backgrounds in North America, the Netherlands, and England. Thus, the type of literacy practices that parents use with preschool children generalize across nationalities, languages, and SES boundaries. The number of children's books owned was comparable with previous studies (Foy & Mann, 2003; Sénéchal et al., 1998). The frequency with which the parents read to their children during preschool (approximately 60% reading daily) was also consistent with previous studies (Sénéchal et al., 1998; Wood, 2002). However, more parents in our study than in previous studies (Haney, & Hill, 2004; Wood, 2002) reported engaging in letter-based teaching activities. This difference may be due to our sample being older and closer to the start of formal schooling (mean age = 5.36 years compared with 4 years in Wood's, 2002, study and 3-5 years in Haney & Hill's, 2004, study). As children near formal schooling, parents may be more likely to increase the frequency of letter teaching practices. There was an indication of this in Sénéchal et al.'s (1998) study, with the average frequency of parental teaching being *sometimes* for kindergarten children and *often* for the children beginning Grade 1. Furthermore, in their conclusions to the Baltimore



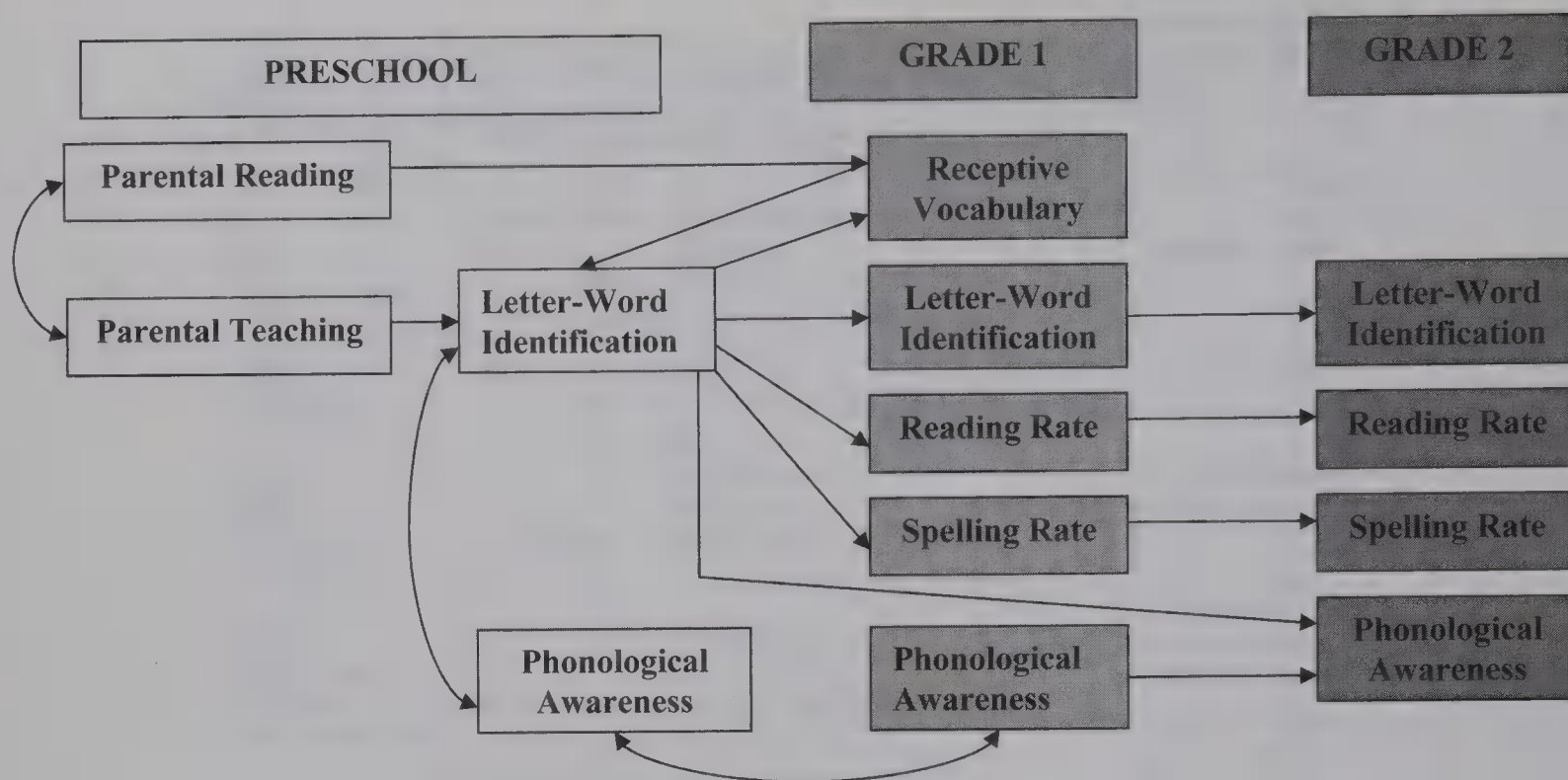


Figure 2. Model of the relationships between early home reading and teaching practices and child literacy and language outcomes, after the effects of age, gender, memory, and nonverbal ability were partialled out. Curved arrows represent paths from Sénéchal and Le Fevre's (2002) model that we did not rigorously test because of the lack of relationships with the focal parent factors but for which we found significant zero-order correlations.

Early Childhood Project, Serpell, Baker, and Sonnenschein (2005) stated that literacy goals were not of paramount importance to parents of prekindergarten-age children.

Consistent with Sénéchal and Le Fevre's (2002; Sénéchal et al., 1998) findings, parental reading and teaching practices loaded on different factors, with little shared variance. Unlike Sénéchal, Le Fevre, and colleagues, we used both Home Literacy Environment Questionnaire items regarding reading frequency and a TRT as our parent-child reading factor. Fritjers et al. (2000) argued that more variance can be accounted for when multiple items are used. Indeed, in post hoc analyses, we found that the correlation between our composite measure and vocabulary (the only significant independent relationship found) was stronger ( $r = .30, p = .001$ ) than was that with the TRT measure alone ( $r = .18, p = .044$ ).

We found a clear difference in the relationships between Parental Reading and Parental Teaching and the literacy and language measures. Consistent with Sénéchal and Le Fevre's (2002) model and other previous studies (e.g., Echols et al., 1996; Nagy & Anderson, 1984; Nagy, Herman, & Anderson, 1985; Scarborough et al., 1991; Sénéchal et al., 1998), parent-child reading was directly related to receptive vocabulary, independently of age, memory, nonverbal ability, and emergent literacy skills. In contrast, Parental Teaching was not directly related to Vocabulary but was directly related to Preschool Letter-Word Identification. The relationship between Parental Teaching and Letter-Word Identification then mediated the relationships between Parental Teaching and all later measures.

We failed to find significant correlations between parent-child reading and children's later reading. Sénéchal and Le Fevre (2002) found that preschool storybook reading was related to Grade 3

reading, but their measure was quite different to ours, incorporating reading vocabulary and reading comprehension. Thus, their result may be largely influenced by a relationship between storybook exposure and reading comprehension that emerges later—an outcome that Leseman and de Jong (1998; de Jong & Leseman, 2001) also found. Furthermore, all of those studies found that relationship was fully mediated by early receptive language skills. Thus, parent-child reading was not directly related to later reading ability.

Studies with older children found that TRT/ART measures were generally related to reading, spelling, phonological processing, and orthographic processing, as well as to vocabulary (Byrne, Fielding-Barnsley, Ashley, & Larsen, 1997; Cipielewski & Stanovich, 1992; Cunningham & Stanovich, 1990, 1991; Echols et al., 1996; McBride-Chang, Manis, Seidenberg, Custodio, & Doi, 1993; Stanovich & West, 1989). It is possible that these different results between samples of beginner and more experienced readers arise because these recognition tests are quite different measures when used with prereaders than when used with older independent readers. With prereaders, parents complete the TRT/ART, so it is a measure of the parent's exposure to children's literature and, thereby, a measure of parental reading to the child. For the pre-reading child, being read to is largely an oral language experience and so the TRT/ART is related to language development. For an older child who is reading independently and completes the TRT themselves, it is a measure of their own independent print exposure. Greater print exposure results in more experience in word and sentence decoding and recognition and in reading comprehension, as well as ongoing exposure to new vocabulary. This explains the relationships between TRT/ART and a range of reading, spelling,

and language measures in older samples that are not found when parent-completed TRT/ART measures are used in younger samples.

Consistent with Sénéchal and Le Fevre's (2002) model and other studies (Baker et al., 1998; Cunningham & Stanovich, 1993; Evans et al., 2000; Sénéchal et al., 1998), neither parent-child reading nor literacy teaching was significantly related to phonological awareness at Preschool or Grade 1. Parental teaching was significantly related to phonological awareness at Grade 2; however, this was completely explained by differences in letter identification at preschool. Foy and Mann (2003) found a weak direct relationship between TRT and rhyme, but not phoneme, awareness. Our composite measure of phonological awareness included both rhyme and phoneme awareness, as did the measure used by Sénéchal, Le Fevre, and colleagues (Sénéchal & Le Fevre, 2002; Sénéchal et al., 1998). This might explain why our studies did not find this relationship. This explanation is further supported by our post hoc analyses that revealed a significant weak correlation ( $r = .19$ ) between our composite reading factor and rhyme awareness at preschool (but not at later phases) that was independent of age, memory, and nonverbal ability. The correlation with phonological segmentation (more focused on phoneme awareness) was not significant.

One reason why stronger relationships with phonological awareness are not found in these studies may also be the parental practices asked about. Our reading questions, like the other studies (Foy & Mann, 2003; Sénéchal & Le Fevre, 2002), focused on story book reading, and our teaching questions focused on letter- or word-based teaching—practices likely to foster vocabulary and letter knowledge, respectively, rather than phonological awareness. Previous studies found that the frequency of rhyming activities (Fernandez-Fein & Baker, 1997) and of reading alphabet books (Murray, Stahl, & Ivey, 1996) was related to children's phonological awareness. Future research needs to canvass a wider range of home literacy practices (including rhyming and letter sound teaching) to clarify whether parental literacy practices can foster phonological awareness.

In contrast to our results with parent-child reading, parental teaching was related to preschool letter identification, independently of associations with age, memory, nonverbal ability, and vocabulary. This relationship then completely mediated the relationships between parental teaching and later word reading and spelling measures, as well as phonological awareness. Theoretical models and empirical evidence supports letter knowledge around school entry as an important predictor of early reading acquisition (Adams, 1995; Scarborough, 1998). Our results show that it is via their relationship with this emergent literacy skill, rather than with early phonological awareness, that home teaching practices (at least as we measured them) are related to Grade 1 word reading and spelling, and that is then directly related to Grade 2 word reading and spelling. The pattern of relationships with spelling was very similar to that with word reading, and the pattern of relationships was similar regardless of whether the outcome was word reading accuracy or rate. Thus, our results confirm that parental teaching practices are significant predictors of emergent literacy skills at the beginning of formal education but that they only have an indirect effect on literacy skills after formal schooling commences. de Jong and Leseman (2001) also found that measures of the early home environment were related to literacy skills at the

start of formal schooling but did not show direct relationships to these skills within 1 year of formal schooling.

We also included a measure of the child's interest in being read to (parent report). Previous studies found that this was related to the frequency of parent-child reading (Crain-Thoreson & Dale, 1992; Olofsson & Niedersoe, 1999) and loaded on the same factor (Dunning et al., 1994). We found child interest was related to both parent-child reading and teaching, having a slightly stronger relationship with parent-child reading. However, unlike previous studies (Fritjers et al., 2000; Olofsson & Niedersoe, 1999; Payne et al., 1994; Scarborough et al., 1991; Sénéchal et al., 1998), we found no relationship between child interest and any of the outcome measures. This might be explained by characteristics of our measure. We had only a single item based on parent report, which was highly skewed (most parents reported their child was very interested).

Our results, combined with previous studies, confirm that similar early home literacy practices are engaged in across a range of SES and cultural backgrounds, and show similar relationships with children's language and literacy outcomes. This is consistent with Bus et al.'s (1995) conclusion that the effects of home literacy practices did not differ among SES; even in low SES families, more shared reading had a positive impact on children's outcomes. However, in the Baltimore Early Childhood Project, Serpell et al. (2005) found that home literacy practices did differ with SES. Therefore, it is a limitation of our study that we did not directly measure SES and compare practices and relationships across SES groups. Notwithstanding that, a major conclusion of the Baltimore Early Childhood Project was that parental practices and beliefs (what they called, the intimate culture) were more important to literacy development than SES. After controlling for parent education, an indicator of SES, Sénéchal and Le Fevre (2002) also found that parent literacy practices were independently related to the language and literacy outcome measures. Burgess (2005) found that a complex network of environmental, educational, and attitudinal variables—not just SES—explained the quality of early home literacy environments provided by teen mothers.

This study adds to the evidence that preschool literacy teaching practices in the home environment are more important than storybook reading in fostering emergent literacy skills and that parents should be encouraged to do more than just read to their children. However, further work is needed to determine the factors that predict parents choosing to engage in more formal literacy teaching over, or in addition to, storybook reading. It may be important for future research to consider this within a developmental context. For example, what specific teaching practices are important for fostering emergent literacy skills at different developmental points prior to school entry, and to what extent are parents sensitive to their child's developing literacy skills, adapting their literacy practices accordingly? Longitudinal studies that begin at an earlier age and that examine a broader range of home literacy practices, ideally with both self-report and observational measures, are needed to clarify developmental changes in the type of practices used and in their relationships to different literacy and language outcomes. We only examined practices in the few months prior to school entry and only examined storybook reading and letter-based teaching activities. Nursery rhymes and other sound games may be more important practices with younger children (e.g., see Bradley



& Bryant, 1983) and may be more strongly related to the other important preliterate skill of phonological awareness.

Parent factors may influence the type of practices engaged in and their effectiveness in promoting children's literacy development. In the current study, we did not determine which parent completed the Home Literacy Environment Questionnaire. It would be of interest in future studies to determine whether fathers and mothers differ in the practices they engage in. Differences in parents' own literacy skills and interest levels may also contribute to differences in the practices engaged in and their effects on child outcomes.

Another factor that may be important, but which is also not addressed in the current study, is the emotional quality of the parent-child literacy interactions (Baker et al., 1998; de Jong & Leseman, 2001; Serpell et al., 2005). de Jong and Leseman (2001) found that the social-emotional quality during parent-child reading was related to vocabulary development and, thereby, to later reading comprehension. Serpell et al. (2005) found that the social-emotional quality was also related to the child's reading motivation and, thereby, to reading frequency once at school. More frequent reading by the child once at school was related to better reading achievement. However, little is currently known about the way social-emotional quality during parental literacy teaching practices is related to children's literacy outcomes. It is possible that unless the parent feels comfortable and competent engaging the child in literacy teaching practices, the social-emotional quality will be less than optimal, and the outcomes may be less satisfactory. This is yet another way in which parents' own literacy skill and interest levels might be related to children's literacy outcomes.

In conclusion, despite our finding that home literacy teaching practices are more important to subsequent literacy development than parent-child reading, reading should not be dismissed in favor of formal teaching practices. Parent-child reading is related to greater vocabulary development, which others have shown is an important predictor of later reading comprehension. Depending on a positive socioemotional context, parent-child reading is also important in fostering the child's own motivation, which has important long-term consequences for their reading as well. Although the direct relationships of parental teaching practices are more obvious earlier in literacy emergence, the indirect relationships of early parent reading may not become apparent until the focus of reading changes from basic word decoding to comprehension, around Grade 3. Thus, parents should be encouraged to engage in both reading and teaching activities to optimize a range of important literacy-related outcomes after school entry.

## References

- Adams, M. (1995). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.
- Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1985). *Becoming a nation of readers: The report of the Commission on Reading*. Champaign, IL: Centre for the Study of Reading.
- Angus and Robertson Bookworld. (1999). *Angus and Robertson's 100 all-time favourite children's books*. Retrieved November 25, 1999, from <http://www.bookworld.com.au/kidstop1002.htm>
- Baker, L., Fernandez-Fein, S., Scher, D., & Williams, H. (1998). Home experiences related to the development of word recognition. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 263-287). Mahwah, NJ: Erlbaum.
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182.
- Bradley, L., & Bryant, P. E. (1983, February 3). Categorizing sounds and learning to read—A causal connection. *Nature*, 301, 419-421.
- Burgess, S. (2005). The preschool home literacy environment provided by teenage mothers. *Early Child Development and Care*, 175, 249-258.
- Burgess, S. R., Hecht, S. A., & Lonigan, C. J. (2002). Relations of the home literacy environment (HLE) to the development of reading-related abilities: A one-year longitudinal study. *Reading Research Quarterly*, 37, 408-426.
- Bus, A. G., van IJzendoorn, M. H., & Pellegrini, A. D. (1995). Joint book reading makes success in learning to read: A meta-analysis of intergenerational transmission of literacy. *Review of Educational Research*, 65, 1-21.
- Byrne, B., Fielding-Barnsley, R., Ashley, L., & Larsen, K. (1997). Assessing the child's and the environments contributions to reading acquisition: What we know and what we don't know. In B. Blachman (Ed.), *Foundations of early reading acquisition and dyslexia: Implications for early intervention* (pp. 265-286). Mahwah, NJ: Erlbaum.
- Cipielewski, J., & Stanovich, K. E. (1992). Predicting growth in reading ability from children's exposure to print. *Journal of Experimental Child Psychology*, 54, 74-89.
- Crain-Thoreson, C., & Dale, P. S. (1992). Do early talkers become early readers? Linguistic precocity, preschool language and early reading. *Developmental Psychology*, 28, 421-429.
- Cunningham, A. E., & Stanovich, K. E. (1990). Assessing print exposure and orthographic processing skill in children: A quick measure of reading experience. *Journal of Educational Psychology*, 82, 733-740.
- Cunningham, A. E., & Stanovich, K. E. (1991). Tracking the unique effects of print exposure in children: Associations with vocabulary, general knowledge, and spelling. *Journal of Educational Psychology*, 83, 264-274.
- Cunningham, A. E., & Stanovich, K. E. (1993). Children's literacy environments and early word recognition subskills. *Reading and Writing: An Interdisciplinary Journal*, 5, 193-204.
- de Jong, P. F., & Leseman, P. P. M. (2001). Lasting effects of home literacy on reading achievement in school. *Journal of School Psychology*, 39, 389-414.
- Dolch, E. W. (1936). A basic sight vocabulary. *Elementary School Journal*, 36, 456-460.
- Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test—Revised*. Circle Pines, MN: American Guidance Service.
- Dunning, D. B., Mason, J. N., & Stewart, J. P. (1994). Reading to preschoolers: A response to Scarborough and Dobrich (1994) and recommendations for future research. *Developmental Review*, 14, 324-339.
- Durkin, D. (1966). *Children who read early*. New York: Teachers College Press.
- Echols, L. D., West, R. F., Stanovich, K. E., & Zehr, K. S. (1996). Using children's literacy activities to predict growth in verbal cognitive skills: A longitudinal investigation. *Journal of Educational Psychology*, 88, 296-304.
- Education Queensland. (2006). *Enrollment statistics*. Retrieved April 20, 2006, from <http://education.qld.gov.au/schools/statistics/pdfs/2004sec1tbl1-3.pdf>
- Elbro, C., Borström, I., & Petersen, D. K. (1998). Predicting dyslexia from kindergarten: The importance of distinctness of phonological representations of lexical items. *Reading Research Quarterly*, 33, 36-60.
- Evans, M. A., Shaw, D., & Bell, M. (2000). Home literacy activities and their influence on early literacy skills. *Canadian Journal of Experimental Psychology*, 54, 65-75.
- Fawcett, A. J., & Nicolson, R. I. (1996). *The Dyslexia Screening Test*. London: Psychological Corporation.

- Fawcett, A. J., Singleton, C. H., & Peer, L. (1998). Advances in early years screening for dyslexia in the United Kingdom. *Annals of Dyslexia*, 48, 57–88.
- Fernandez-Fein, S., & Baker, L. (1997). Rhyme and alliteration sensitivity and relevant experiences among preschoolers from diverse backgrounds. *Journal of Literacy Research*, 29, 433–459.
- Foy, J. G., & Mann, V. (2003). Home literacy environment and phonological awareness in preschool children: Differential effects for rhyme and phoneme awareness. *Applied Psycholinguistics*, 24, 59–88.
- Fritjers, J. C., Barron, R. W., & Brunello, M. (2000). Direct and mediated effects of home literacy and literacy interest on prereaders oral vocabulary and early written language skill. *Journal of Educational Psychology*, 92, 466–477.
- Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall International.
- Haney, M., & Hill, J. (2004). Relationships between parent-teaching activities and emergent literacy in preschool children. *Early Child Development and Care*, 174, 215–228.
- Jackson, E. N., Donaldson, G. W., & Cleland, L. N. (1988). The structure of precocious reading ability. *Journal of Educational Psychology*, 80, 234–243.
- Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *Handbook of social psychology* (4th ed., Vol. 1, pp. 233–265). New York: McGraw-Hill.
- Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- Leseman, P. P. M., & de Jong, P. F. (1998). Home literacy: Opportunity, instruction, cooperation, and social-emotional quality predict early reading achievement. *Reading Research Quarterly*, 33, 294–318.
- Mann, V., & Liberman, I. Y. (1984). Phonological awareness and verbal short-term memory: Can they presage early reading success? *Journal of Learning Disabilities*, 17, 592–599.
- McBride-Chang, C., Manis, F. R., Seidenberg, M. S., Custodio, R. G., & Doi, L. M. (1993). Print exposure as a predictor of word reading and reading comprehension in disabled and normal readers. *Journal of Educational Psychology*, 85, 230–238.
- Meyer, L. A., Wardrop, J. L., Stahl, S. A., & Linn, R. L. (1994). Effects of reading storybooks aloud to children. *Journal of Educational Research*, 88, 69–85.
- Molfese, V. J., Molfese, D. L., & Modgline, A. A. (2001). Newborn and preschool predictors of second-grade reading scores: An evaluation of categorical and continuous scores. *Journal of Learning Disabilities*, 34, 545–554.
- Murray, B. A., Stahl, S. A., & Ivey, M. G. (1996). Developing phoneme awareness through alphabet books. *Reading and Writing: An Interdisciplinary Journal*, 8, 307–322.
- Nagy, W. E., & Anderson, R. C. (1984). The performance of learning (reading) disabled children on a test of spoken language. *Reading Research Quarterly*, 19, 304–330.
- Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly*, 20, 233–253.
- Nicolson, R. I., & Fawcett, A. J. (1996). *The Dyslexia Early Screening Test*. London: Psychological Corporation.
- Olofsson, A., & Niedersoe, J. (1999). Early language development and kindergarten phonological awareness as predictors of reading problems: From 3 to 11 years of age. *Journal of Learning Disabilities*, 32, 464–472.
- Olson, R. K., Wise, B., Johnson, M. C., & Ring, J. (1997). The etiology and remediation of phonologically based word recognition and spelling disabilities: Are phonological deficits the “hole” story? In B. A. Blachman (Ed.), *Foundations of reading acquisition and dyslexia: Implications for early intervention* (pp. 305–326). Mahwah, NJ: Erlbaum.
- Payne, A. C., Whitehurst, G. J., & Angell, A. L. (1994). The role of home literacy environment in the development of language ability in preschool children from low-income families. *Early Childhood Research Quarterly*, 9, 427–440.
- Preacher, K. J., & Leonardelli, G. J. (2001, March). Calculation for the Sobel Test: An interactive calculation tool for mediation tests [Computer software]. Retrieved March 1, 2007, from <http://www.quantpsy.org>
- Raven, J. C., Court, J. H., & Raven, J. (1986). *Raven's Coloured Matrices*. London: H. K. Lewis.
- Scarborough, H. S. (1998). Early identification of children at risk for reading disabilities: Phonological awareness and some other promising predictors. In B. K. Shapiro, P. J. Accardo, & A. J. Capute (Eds.), *Specific reading disability: A view of the spectrum* (pp. 75–119). Timonium, MD: York Press.
- Scarborough, H. S., Dobrich, W., & Hager, M. (1991). Preschool literacy experience and later reading achievement. *Journal of Learning Disabilities*, 24, 508–511.
- Sénéchal, M., & Le Fevre, J. (2002). Parental involvement in the development of children's reading skill: A five-year longitudinal study. *Child Development*, 73, 445–460.
- Sénéchal, M., Le Fevre, J., Hudson, E., & Lawson, P. (1996). Knowledge of storybooks as a predictor of young children's vocabulary. *Journal of Educational Psychology*, 88, 520–536.
- Sénéchal, M., Le Fevre, J., Thomas, E. M., & Daley, K. E. (1998). Differential effects of home literacy experiences on the development of oral and written language. *Reading Research Quarterly*, 33, 96–116.
- Serpell, R., Baker, L., & Sonnenschein, S. (2005). *Becoming literate in the city: The Baltimore Early Childhood Project*. Cambridge, England: Cambridge University Press.
- Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and non-experimental studies: New procedures and recommendations. *Psychological Methods*, 7, 422–445.
- Singleton, C. H., Thomas, K. V., & Leedale, R. C. (1997). *CoPS 1 Cognitive Profiling System: Windows edition*. Beverley, East Yorks, United Kingdom: Lucid Research Limited.
- Sobell, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 290–312). Englewood Cliffs, NJ: Prentice Hall.
- Stainthorp, R., & Hughes, D. (2000). Family literacy activities in the homes of successful young readers. *Journal of Research in Reading*, 23, 41–54.
- Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, 24, 402–433.
- Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin*, 101, 192–212.
- Wood, C. (2002). Parent-child pre-school activities can affect the development of literacy skills. *Journal of Research in Reading*, 25, 241–258.
- Woodcock, R. W. (1997). *Woodcock Diagnostic Reading Battery*. Itasca, IL: Riverside Publishing.

(Appendixes follow)



## Appendix A

Percentage of Parents ( $N = 143$ ) Indicating Recognition for Book and Foil Titles on the TRT

Real Title	Percentage recognized	Foils	Percentage recognized
Green Eggs and Ham	53.5	How Andrew Saved the Day	3.5
Corduroy	10.6	Toby the Terrible Tip Truck	3.5
Winnie the Pooh	83.1	Are You My Father?	9.9
Possum Magic	28.9	Postman Pat at the Beach	19.0
Tooth Fairy	22.5	Old Fox	2.8
The Very Hungry Caterpillar	68.3	Hello Morning, Hello Day	0.7
Are You My Mother?	38.7	Dairy Wood	2.1
The Velveteen Rabbit	18.3	The Very Naughty Fairy	2.1
The Cat in the Hat	68.3	Elephant Magic	2.8
Mike Mulligan and His Steam Shovel	4.9	Thomas the Tank Engine's White Christmas	19.7
Koala Lou	9.2		
Where's Spot?	76.8		
The Complete Adventures of Blinky Bill	33.8		
Hairy MacLary From Donaldson's Dairy	54.2		
Where the Wild Things Are	18.3		
Who Sank the Boat?	12.0		
Harry the Dirty Dog	38.0		
We're Going on a Bear Hunt	38.0		
Saggy Baggy Elephant	22.5		
Just Me and My Dad	8.5		

Note. TRT = Title Recognition Test.

## Appendix B

## Home Literacy Environment Questions and Frequencies of Responses

1. In a typical week, how often do you, or other members of the family, read to your child?

At bedtime:	<input type="checkbox"/> never	8.4%
	<input type="checkbox"/> Once	9.1%
	<input type="checkbox"/> 2 times	11.9%
	<input type="checkbox"/> 3 times	10.5%
	<input type="checkbox"/> 4 times	14.7%
	<input type="checkbox"/> 5 times	17.5%
	<input type="checkbox"/> 6 times	8.4%
	<input type="checkbox"/> 7 times	19.6%

## At other times:

<input type="checkbox"/> never	2.8%
<input type="checkbox"/> Once	12.6%
<input type="checkbox"/> 2 times	18.9%
<input type="checkbox"/> 3 times	28.0%
<input type="checkbox"/> 4 times	14.7%
<input type="checkbox"/> 5 times	10.5%
<input type="checkbox"/> 6 times	1.4%
<input type="checkbox"/> 7 times	5.6%
<input type="checkbox"/> more often. How often? _____	5.6%

(Appendixes continue)



2. Please estimate the number of children’s books per child available in your household?

<input type="checkbox"/>	1- 5	2.1%	<input type="checkbox"/>	41 – 50	15.3%
<input type="checkbox"/>	6 – 10	3.5%	<input type="checkbox"/>	51 – 60	8.3%
<input type="checkbox"/>	11 – 15	16.0%	<input type="checkbox"/>	61 – 70	1.4%
<input type="checkbox"/>	16 – 20	8.3%	<input type="checkbox"/>	71 – 80	0.7%
<input type="checkbox"/>	21 – 25	9.7%	<input type="checkbox"/>	81 – 90	0.0%
<input type="checkbox"/>	26 – 30	4.2%	<input type="checkbox"/>	91 – 100+	13.9%
<input type="checkbox"/>	31 – 40	16.7%			

3. When being read a story, how interested does your child appear to be?

<input type="checkbox"/>	Not interested at all	0.7%
<input type="checkbox"/>	Slightly interested	6.3%
<input type="checkbox"/>	Quite interested	25.2%
<input type="checkbox"/>	Very interested	67.8%
<input type="checkbox"/>	Don’t know	0.0%

4. Please indicate how often, on average, your child would normally engage in the following activities (filler items not reported here). Please circle the appropriate number, where 1 means *daily* and 6 means *never*.

	<i>Daily</i>	<i>Weekly</i>	<i>Fortnightly</i>	<i>Monthly</i>	<i>Less than Monthly</i>	<i>Never</i>
▪ Go to Library	1	2	3	4	5	6
Percent Responding:	0.0	35.7	11.9	15.4	16.1	21.0

5. In a typical week, how often do you, or another family member, engage in the following activities (filler items not reported here)? Please circle a number from 1 – 5 where 1 means *never* and 5 means *very often*. If you don't teach an activity because your child knows how to do it already, circle 6 = NA for *not applicable*

*Never      Rarely      Sometimes      Often      Very often      NA*

I/we teach my child:

(Percent responding)

▪ the alphabet letters	0	1.4	13.3	33.6	47.6	4.2
▪ how to write own name	0	0.7	14.0	31.5	44.8	9.1
▪ how to read words	2.1	5.6	35.0	28.0	29.4	0.0

### Appendix C

#### Reading and Spelling Rate Word Lists

the	this	we	new	before	get
of	had	him	some	must	here
and	not	been	could	where	both
to	are	has	these	much	under
a	but	when	two	your	never
in	from	who	many	well	know
that	or	will	then	down	us
is	have	no	do	because	old
was	an	if	first	just	hurt
he	they	out	any	those	wash
for	which	so	my	how	thank
it	one	said	now	too	sing
with	you	what	like	little	fly
as	were	up	our	good	laugh
his	her	its	over	very	jump
on	all	about	me	make	ate
be	she	into	made	own	more
at	there	them	after	see	than
by	would	can	did	work	other
I	their	only	many	long	time

Received October 14, 2005

Revision received October 23, 2007

Accepted December 13, 2007 ■



# Repeated Reading Intervention: Outcomes and Interactions With Readers' Skills and Classroom Instruction

Patricia F. Vadasy and Elizabeth A. Sanders  
Washington Research Institute

This study examined effects of a repeated reading intervention, *Quick Reads*, with incidental word-level scaffolding instruction. Second- and third-grade students with passage-reading fluency performance between the 10th and 60th percentiles were randomly assigned to dyads, which were in turn randomly assigned to treatment (paired tutoring,  $n = 82$ ) or control (no tutoring,  $n = 80$ ) conditions. Paraeducators tutored dyads for 30 min per day, 4 days per week, for 15 weeks (November–March). At midintervention, most teachers with students in the study were formally observed during their literacy blocks. Multilevel modeling was used to test for direct treatment effects on pretest–posttest gains as well as to test for unique treatment effects after classroom oral text reading time, 2 pretests, and corresponding interactions were accounted for. Model results revealed both direct and unique treatment effects on gains in word reading and fluency. Moreover, complex interactions between group, oral text reading time, and pretests were also detected, suggesting that pretest skills should be taken into account when considering repeated reading instruction for 2nd and 3rd graders with low to average passage-reading fluency.

**Keywords:** multilevel modeling, fluency, repeated reading, verbal efficiency, paraeducators

Repeated reading is a widely used practice (Meyer & Felton, 1999; Samuels, 1979) to increase reading fluency (National Reading Panel, 2000); yet, despite its use and acceptance, it is not an entirely standardized procedure. For example, there is no consensus regarding how long and intense it must be to be effective (Chard, Vaughn, & Tyler, 2002), and most research on repeated reading has been conducted on relatively short interventions (Wolf & Katzir-Cohen, 2001). Repeated reading interventions may require students to reread the same texts or different texts, and again, it is not clear which of these methods is more beneficial (Kuhn & Stahl, 2003). When repeated reading is implemented, it may be assisted by an adult reader or teacher, and it is unclear whether the most effective method depends upon the student's skills and developmental needs (National Reading Panel, 2000). Finally, the type of text that is most beneficial to use in a repeated reading intervention is not well specified (Kuhn & Stahl, 2003). There are equivocal findings whether independent or instructional-level text

is associated with greater fluency gains (Kuhn & Stahl, 2003). Less is understood about the role of text characteristics that, as Menon and Hiebert (2005) suggested, may mediate instructional and fluency gains. These text characteristics include total number of words, word repetitions, and word frequency (i.e., the likelihood that the words reread will be encountered and reinforced in other grade-level texts or instructional content). The number of word repetitions or exposures during reading instruction may contribute to the growth of word recognition skills (Juel & Roper-Schneider, 1985).

Researchers' increased understanding of the subtypes of dysfluent students (Berninger, Abbott, Billingsley, & Nagy, 2001) and stages of fluency development raises questions about the grade levels and developmental stages when repeated reading may be most effective for fluency as well as comprehension. The importance of specifying treatment effects for particular stages of fluency development is underscored by the work of Schwanenflugel et al. (2006). In support of their simple reading fluency model they found that for children in Grades 1 to 3, a single reading fluency factor that included word reading accuracy and efficiency skills and object naming was related to comprehension. The National Reading Panel (2000) summarized limitations and directions for future research on fluency, including the need for explicit treatment descriptions of amount of rereading, type of feedback, and difficulty level of materials. The present study attempts to specify more fully the features of a supplemental fluency intervention for students with a range of emerging fluency performance. We also examine features of students' classroom instruction that may influence reading skill growth and interact with fluency treatment.

Fluent reading reflects the development and consolidation of subskills that permit the student to allocate attention to constructing a meaning of the text (LaBerge & Samuels, 1974). According to Perfetti's (1985, 1992; Perfetti & McCutchen, 1987) verbal efficiency theory, the efficiency of phonological, orthographic, and

---

Patricia F. Vadasy and Elizabeth A. Sanders, Washington Research Institute, Seattle, Washington.

Grant R305G040103A from the Institute of Education Sciences, U.S. Department of Education, supported this research. We especially acknowledge Sueanne Sluis for coordinating the intervention and the classroom observations. We also thank Sarah Tudor and Kathryn Compton for coordinating student assessments; and we thank coaches, testers, teacher observers, tutors, and data entry staff for their work on the project, including Sarah Barton, Katy Compton, Alison Fiorito, Christine Haas, Rayma Haas, Robin Horton, Gwenneth Kunde, Ruthanne McPhaden, Siobhain Mogenssen, Devon Parris, Lindsay Schwartz, Laura Root, Tyler Rothnie, Laurel Rustmeyer, and Lynn Youngblood. Last but not least, we are most grateful to the teachers, staff, and children of the schools for their support and participation in this study.

Correspondence concerning this article should be addressed to Patricia F. Vadasy, Washington Research Institute, 150 Nickerson Street, Suite 305, Seattle, WA 98109. E-mail: pvadasy@wri-edu.org

semantic processes underlies word recognition and retrieval, and differences in these lexical retrieval processes influence comprehension. Slow and inefficient word identification creates a bottleneck that diverts cognitive resources required for comprehension. Verbal efficiency theory predicts that efficiency in word and text reading accuracy and speed will produce improved comprehension. The research on training in verbal efficiency, however, is mixed. Two types of training have been used to develop fluency: training to recognize single words and practice in reading connected text (repeated reading). Fleisher, Jenkins, and Pany (1979) trained students to read words in isolation and in word phrases, and although both types of training increased decoding accuracy and speed for single words, training did not improve comprehension. Others have also failed to show transfer to reading comprehension through training in single-word reading (Yuill & Oakhill, 1988). In a replication of the Fleisher et al. study, Tan and Nicholson (1997) used training in single-word reading to produce gains in comprehension. Others have reported training in single words or connected text generalized to comprehension (Levy, Abello, & Lysynchuk, 1997; O'Shea, Sindelar, & O'Shea, 1985; Samuels, Dahl, & Archwamety, 1974), with reports of mixed effects (Roth & Beck, 1987). Findings on verbal efficiency training are difficult to interpret due to differences in student age and developmental level, training focus, and comprehension tasks.

For practical purposes we define fluency as "rate and accuracy in oral reading" (Torgesen, Rashotte, & Alexander, 2001). However, our assessments reflect an emerging understanding of fluency that encompasses subcomponent skills and executive coordination processes (Berninger et al., 2001) as well as the developmental nature of fluency (Kame'enui, Simmons, Good, & Harn, 2001). Of particular interest in this study of second- and third-grade students was word reading efficiency or automaticity, which is also represented in Schwanenflugel et al.'s (2006) simple reading fluency model of emergent fluency development. Both the more classic and contemporary theories of fluency predict that word recognition and retrieval skills explain individual variance in children's fluency rates in reading connected text.

Several published reviews of fluency instruction have summarized the theory and nature of fluency interventions (Wolf & Katzir-Cohen, 2001), features and research on fluency practices (Kuhn & Stahl, 2003), and the findings on fluency interventions for children with learning disabilities (Chard et al., 2002). In their review of the history of fluency interventions and their components, Wolf and Katzir-Cohen described a theory-based, multi-component fluency intervention based on information processing models (LaBerge & Samuels, 1974; Perfetti, 1985) that predict that fast acting subskills enable students to allocate sufficient attention to higher level text comprehension. Wolf and Katzir-Cohen described more recent research that captures both the developmental nature of fluency (Kame'enui et al., 2001) and multiple processing systems (phonological, orthographic, and morphological) that contribute to text-level reading fluency. This review highlighted several important limitations. First, most interventions were relatively short—between 1 and 15 days in length. More recent research has underscored the difficulty of remediating students with poor fluency (i.e., Torgesen et al., 2001) and the need for intensive intervention. Second, most of the studies in this review did not include a control group. Third, comprehension was assessed inconsistently. And fourth, studies did not examine lexical-level rate

change that may be most directly influenced by repeated reading (Faulkner & Levy, 1999; Torgesen et al., 2001).

Kuhn and Stahl (2003) identified 15 studies of repeated reading that included a control group. They found that repeated reading intervention was more effective than the control condition in six studies and in one study only when familiar passages were used. Considering findings of effectiveness based on individual study comparisons (adding weight to studies with multiple comparisons), they found that repeated reading interventions were significantly more effective than control conditions in 8 comparisons and not more effective in 21 comparisons. Differences in types of control groups used in these studies (no treatment or nonrepeated reading control) made it difficult to draw firm conclusions across these studies. Kuhn and Stahl found some advantage for reading more difficult text and that comprehension gains were difficult to detect using standardized comprehension measures (rather than cloze tasks).

Finally, Chard et al. (2002) reviewed 24 studies of fluency interventions provided for elementary students with learning disabilities. They found that repeated reading with a model was most effective, particularly for students with low fluency. A teacher or adult model was more helpful than a tape or computer model. Fluency growth was associated with comprehension growth in many of the studies. Although findings did not suggest an optimal number of rereadings, the authors concluded that effective components of fluency intervention for students with learning disabilities include rereading difficult text with adult feedback and corrections. The study design we describe here addresses limitations in research designs found in these earlier reviews of fluency interventions, including lack of a control group, low intensity of treatment, and lack of standardized word reading efficiency and comprehension measures.

### Intervention Research on Supplemental Instruction

Supplemental reading interventions play an important role in many children's reading instruction. They may be provided by teachers or by paraeducators, and in individual or small-group, pull-out sessions. A small number of these interventions, particularly early phonological awareness and phonics-based programs, have been evaluated in rigorous research designs (*Road to the Code*: Blachman, Ball, Black, & Tangel, 1994; Blachman, Tangel, Ball, Black, & McGraw, 1999; *Ladders to Literacy*: Fuchs et al., 2001, 2002; *Peer Assisted Learning Strategies*: Fuchs et al., 2001, 2002; *Spell Read P.A.T.*: Rashotte, MacPhee, & Torgesen, 2001). There is limited research on supplemental instruction using programs that focus on fluency skills (*Read Naturally*: Hasbrouck, Ihnot, & Rogers, 1999; *Great Leaps*: Mercer, Campbell, Miller, Mercer, & Lane, 2000). Because the fluency intervention in the present study was used to supplement classroom instruction and was implemented by paraeducator tutors, we consider features associated with effective supplemental tutoring interventions. Juel (1996) summarized reading activities significantly associated with literacy growth in tutoring programs, including direct letter-sound instruction and the introduction of high-frequency and pattern words. Wasik's (1998b) review of 17 studies of volunteer tutoring in reading skills identified four features common to programs: a designated coordinator with a background in reading instruction; structured tutoring sessions that included reading new material,



reading familiar books, and focusing on word analysis and alphabets; writing composition; and tutor training. Wasik (1998a) and others (Jenkins & Jenkins, 1987; Reisner, Petry, & Armitage, 1990; Venezky & Jain, 1996) also recommended that tutoring be coordinated with classroom instruction, although well-designed studies on this issue are lacking. The fluency intervention we describe includes many features identified in effective tutoring programs: a structured format, a focus on alphabets and high-frequency words, tutor training, and tutor scaffolding and modeling.

Research on paraeducator-supplemented instruction provides another source of information on effective tutoring. It includes studies of kindergarten and first-grade literacy interventions implemented by instructional assistants (Blachman et al., 1994; Gunn, Biglan, Smolkowski, & Ary, 2000; Gunn, Smolkowski, Biglan, & Black, 2002; Gunn, Smolkowski, Biglan, Black, & Blair, 2005; Simmons, Kame'enui, Stoolmiller, Coyne, & Harn, 2003; Torgesen, Wagner, Rashotte, et al., 1999; Vadasy, Jenkins, & Pool, 2000; Vadasy, Sanders, & Abbott, in press; Vadasy, Sanders, Jenkins, & Peyton, 2002; Vadasy, Sanders, & Peyton, 2006). Critical features of effective paraeducator-implemented instruction include training and research-based standard treatment protocols that paraeducators can learn to use with fidelity.

### Research Questions

This study was designed to assess effects of a repeated reading intervention, *Quick Reads* (Hiebert, 2003), with incidental word-level scaffolding instruction on student gains as compared to regular classroom reading instruction. Research questions were the following:

1. What are the direct effects of this intervention, provided by paraeducators to pairs of second- and third-grade students with low to average passage-reading fluency (PRF), on students' word reading, fluency, and comprehension gains?
2. What are the unique effects of this intervention after accounting for classroom oral text reading (OTR) time, pretest rapid automatized naming (RAN), and pretest word reading accuracy? Are treatment effects moderated by these variables?

We expected the greatest reading fluency gains to occur for students with stronger initial word reading accuracy, as well as for students who received more classroom oral reading practice (relative to their peers). Additional oral reading practice might boost fluency development; although as Berninger et al. (2001) suggested, subtypes of students with poor fluency may require differential treatments to address specific problems in lexical retrieval and coordination of language processes. We were unable to identify subtypes of fluency deficits in this study, and we assumed that classroom oral reading practice opportunities would generally benefit fluency outcomes. We therefore collected information on text reading opportunities provided during classroom reading instruction for both treatment and control students, hypothesizing that these activities might benefit both groups.

For descriptive purposes, we also recorded classroom instructional time allocated to word identification and comprehension

skills, as well as organizational features of classroom reading instruction, such as grouping and use of time (Cameron, Connor, & Morrison, 2005). Findings from recent studies using classroom observational systems (Connor, Morrison, & Katch, 2004; Connor, Morrison, & Petrella, 2004; Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998; Taylor, Pearson, Peterson, & Rodriguez, 2005) have suggested that reading instruction varies along several dimensions across classrooms, including grouping strategies, time spent on salient instructional activities, and the interaction of classroom instruction and student skill levels. As Smith and Siegel (2006) noted, attention to routine practice conditions may inform how to best scale up effective reading interventions. In this study paraeducators rather than researchers were used to implement instruction, and a high fidelity of implementation was sought.

### Method

#### Participants

*Student referral and screening.* In the fall of the academic year, second- and third-grade teachers from 13 urban, public elementary schools were asked to refer students who (a) had never been retained, (b) had low rates of reading fluency or comprehension, and (c) would benefit from a fluency intervention (i.e., students with adequate word reading skills to benefit from fluency instruction). Once active parent consents were obtained, referred students were screened for eligibility using two grade-level passages from the Oral Reading Fluency subtest of the Dynamic Indicators of Basic Early Literacy Skills (Good & Kaminski, 2002). Students whose performance was in the 10th to 60th percentile range (approximately averaging 34th percentile) for their grade level on the mean of the two passages were considered eligible for participation (60th percentile was approximated by solving for the respective fluency rate value corresponding to  $Z = 0.25$ ; Hasbrouck & Tindal, 2006). This performance range was specifically 11 to 62 words correct per minute (wcpm) for second graders and 21 to 82 wcpm for third graders. Teachers referred students with a wide range of fluency for study participation. It should be noted that by Grades 2 to 3, reading rate becomes an important reading target in schools. As a result, many students had low rates of reading fluency and below-grade-level decoding skills.

*Group assignment.* Group assignment was a two-stage process. First, eligible students were randomly assigned to dyads (pairs of students) within grade and school using a random number generator. Although it would have been preferable to randomly assign students to dyads within classrooms in order to minimize extreme mismatches between students within a dyad, there were too few eligible students within classrooms to make this practical. Nevertheless, there were sufficient numbers of eligible students for random assignment of dyads within grades (within schools). For schools with uneven numbers of students within grades, we used random selection to identify singletons that were subsequently excluded from participation. Dyads were then randomly assigned to one of two conditions: treatment (supplemental fluency tutoring instruction) or control (no tutoring; classroom instruction only).

*Attrition.* After group assignment, the sample comprised 96 students (48 dyads) in the treatment condition and 92 students (46

dyads) in the control condition. By the end of the study, 14 treatment and 12 control students were lost to attrition (14%). To ensure an operational definition of treatment as well as an unbiased attrition pattern across treatment and control groups, we removed any dyad member from study participation if the corresponding member moved from the school. (Although control pairs did not necessarily receive reading instruction together, we assumed that control pairs shared common reading curricula due to the within-grade, within-school pairings.) The sample used for analyses thus comprised 82 treatment students (41 dyads) and 80 control students (40 dyads). As reported in Table 1, there were no significant differences between groups on grade or status variable frequencies (all  $ps > .05$ ).

**Tutors.** A total of 22 tutors were recruited from their school communities. Tutors' educational levels, general tutoring experience, and experience working with second and third graders varied. All but one had already been working at the school as either an instructional assistant or tutor. Prior to the study, tutors had a range of 0 to 11 years of reading tutoring experience ( $M = 3.6$  years,  $SD = 3.43$ ), and all tutors had at least a high school diploma (15 had bachelor's degrees, 5 with teaching certificates). The average educational attainment of tutors in this study matched the paraeducator competency requirements under the No Child Left Behind Act of 2001. Most tutors (91%) were female, and six (27%) were of minority heritage.

### Intervention

Treatment students received supplemental tutoring in dyads for 30 min per day, 4 days per week, for 15 weeks (November to March) using the *Quick Reads* (Hiebert, 2003) fluency program. Students assigned to the control group received regular classroom instruction, whereas treatment students received tutoring in their respective pairs. Classroom teachers reported that most treatment students ( $n = 58$ , or 70%) missed some portion of classroom reading instruction. For 15 students (18%), the *Quick Reads* intervention served as added reading instruction time: These students typically missed the nonreading activities and instruction such as chess club, computer lab, math, physical education, science, and

social studies. Teachers of 9 students did not report the instruction missed during tutoring.

The *Quick Reads* program includes short, nonfiction passages written for grade levels 1 to 6. Each grade level includes nine topics based on national and state curriculum standards for science and social science. Five passages are written to develop knowledge about each topic. For example, for Level B (second grade), science topics include "Water and Us," "Weather," and "Rocks." Level B social studies topics include "Maps," "Money," and "Jobs Around Us." An important feature of the *Quick Reads* program is its attention to text features to build fluency and comprehension. The number of unique or different words is minimized, and 98% of words used in the texts are high-frequency words or words that reflect grade-level phonics and syllable patterns. Repeated word exposures are designed to build sight word learning. Rare words and singletons are minimized, as students are unlikely to acquire these words as sight words from single exposures and are unlikely to practice them in classroom texts. These text characteristics are hypothesized to develop underlying lexical accuracy and automaticity skills often overlooked in traditional repeated reading remedial interventions.

*Quick Reads* is designed for classroom or small-group use, either in regular reading instruction or as a supplemental intervention. The number of passages students cover daily and the length of the intervention depend on how it is used. If a teacher follows the recommended classroom instructional routine, students use *Quick Reads* for 15 min a day for one semester, or 18 weeks. Each passage is read three times:

1. First read. The teacher activates background knowledge about the topic and asks students to find two words that are challenging. Students read the passage aloud or silently and then write notes or phrases of key ideas.
2. Second read. The teacher reads aloud with the students, setting a model for fluent reading, all reading aloud at the target rate of 1 min. The teacher asks students to "tell the one thing the author wants you to remember."
3. Third read. The teacher instructs students to read as much of the passage as they can read in 1 min. The students then read silently for 1 min, and when the time is up each student records the number of words read. The teacher and student review two comprehension questions together.

The teacher's manual from the *Quick Reads* program outlines placement procedures for students with passage-reading performance of (a) at least 90% accuracy and at least a rate of 50 words per minute (wpm), (b) less than 90% accuracy or a rate of less than 50 wpm, and (c) less than 90% accuracy and a rate of less than 50 wpm. For students in the last category, the manual suggests a modified instructional routine that might include instruction in word reading skills. The manual also suggests teaching options to develop vocabulary, word identification, and comprehension skills. For example, activities are suggested for teaching vocabulary or identifying the main idea of the passage.

**Tutoring sessions.** Rather than expect tutors to choose an appropriate activity to add to the repeated reading steps, we created

Table 1  
Student Characteristics

Characteristic	Treatment		Control		$\chi^2(1, N = 162)$
	N	%	N	%	
Grade 2	58	70.7	52	65.0	0.61
Male	47	57.3	51	63.8	0.70
Minority	56	68.3	57	71.3	0.17
Asian	16	19.5	10	12.5	
Black	22	26.8	23	28.8	
Hispanic	15	18.3	22	27.5	
Mixed/other	3	3.7	2	2.5	
English-language learner	19	23.2	18	22.5	0.01
Special education	13	15.9	15	18.8	0.24
Title 1	63	76.8	58	72.5	0.40

Note. Treatment group ( $N = 82$ ), control group ( $N = 80$ ). Chi-square tests of independence were used to compare categorical frequencies between groups. All  $ps > .05$ .



an instructional package similar to what is suggested in the *Quick Reads* manual. The teacher manual is written for classroom or small-group instruction and assumes that the teacher chooses the enrichment activities, coordinated with student need and other reading instruction. We added brief word-level instruction to meet the needs of the many students with low word reading skills who were referred for study participation, and we allowed tutors up to 5 min per session for this instruction as needed. (We also measured the amount of phonics/word study instruction provided in the classroom to understand better how the intervention coordinated with classroom instruction for low decoders in particular.) All instruction for this intervention was scripted, as described here, to ensure instructional consistency. Each tutoring session had six steps, as follows:

1. Letter/sound practice. Previous experience with an earlier cohort of *Quick Reads* students (Vadasy & Sanders, in press) led us to expect that second- and third-grade students with poor fluency would benefit from brief instruction and review in the alphabetic principle, particularly two-letter spelling patterns, as well as supportive correction strategies for their word miscues. Although this extension activity is not specifically suggested in the teacher's manual, we believed it would strengthen the program for the less skilled students in this study, and it was brief enough not to weaken the recommended intensity of the *Quick Reads* instruction (recommended for 12- to 15-min sessions to cover one passage). We therefore designed and scripted an extension activity for this *Quick Reads* field test that allowed tutors to review letter sounds if needed and provide scaffolded assistance in decoding difficult words. For up to 5 min of each session, tutors used a set of letter sound cards to review and practice accurate and automatic letter-sound correspondences, including vowel teams and consonant blends. Tutors were instructed to provide this practice only for letters and sounds that were problematic for students. If the student could not yet identify the sounds, the tutor directed the student to point to the letters on the cards and say the letter name, picture cue, and letter sound. For letters the student could not automatically match with a sound, the tutor pointed to the letters and asked the student to say the sounds. Time for review was allotted to each student as needed.

Tutors were also trained to provide scaffolding for word reading attempts with graduated levels of assistance to encourage students to apply their alphabetic and decoding skills. Corrective and supportive strategies the tutors used included (a) referring to the letter sound card to retrieve a letter sound, (b) encouraging the student to stretch out the sounds and then say the word fast, (c) assisting the student in segmenting a multisyllable word and then putting the parts together, and (d) telling the student an irregular nondecodable word and having the student reread the word.

2. First passage reading. Tutors introduced the main idea for the first passage, and students skimmed the passage to find difficult words to practice before passage reading. Students then took turns reading and following along with their finger.

3. Second and third passage reading. Both the tutor and students read the passage aloud together twice, with the tutor modeling smooth, accurate, and fluent reading.

4. Fourth passage reading. Each student completed a 1-min timed reading for which the tutor recorded the students' reading rates and accuracies.

5. Comprehension. Tutors and students discussed two comprehension questions.

6. Reading of new passage/rereading of previous passage(s). As time permitted, students reread the previous passage or began a new passage (following the first five steps outlined previously). Students completed Steps 1 through 5 for at least two passages per session.

*Quick Reads placement and coverage.* *Quick Reads* passages are organized by grade level (A = first grade, through F = sixth grade). Each *Quick Reads* level has three books, and each book contains six content areas with five passages per content area, for a total of 90 passages per level. We placed dyads into levels based upon the grade level for which their average pretest reading fluency most closely matched 50th percentile (Hasbrouck & Tindal, 2006). Our sample, after attrition, included 29 dyads placed into Level B and 12 dyads placed into Level C. One dyad originally placed in Level B was moved back to Level A, but all other dyads moved forward to Levels C or D after completing the previous level.

For each student's session, tutors recorded attendance, including the *Quick Reads* passage(s) covered and the fluency rate and accuracy for each timed reading. Treatment students completed an average of 95 passages ( $SD = 18.2$ ), attended an average of 50 tutoring sessions ( $SD = 5.8$ ) or 25 hr of intervention, and averaged 1.9 passages read per session ( $SD = 0.23$ ; range = 1.5–2.9). To determine the difficulty level of the passages students read, we computed the accuracy rate across all attended sessions on *Quick Reads* levels. On average, students read at 95% accuracy within Level A and at 97% accuracy within Levels B, C, and D. These estimates suggest that the students read relatively difficult text at an instructional level for which students needed assistance. More difficult repeated reading materials may be associated with greater fluency improvements (Kuhn & Stahl, 2003).

*Tutor training.* Tutors participated in one initial 4-hr training session by project staff. Training included an overview of reading fluency development and the repeated reading method. Research staff then modeled the use of *Quick Reads* materials and demonstrated procedures for adding instruction/scaffolding in decoding. The tutors practiced the protocols during training and received immediate feedback. Following this training, coaches visited tutors biweekly to provide follow-up training and modeling and to collect data on protocol fidelity.

*Tutor coaching.* Throughout the 15-week intervention, research staff supported and conducted observations of the tutors. Researcher coaches were assigned to specific tutors and conducted a minimum of six observations on each tutor (at least two observations per dyad). Coaches met monthly to discuss tutoring implementation progress.

*Tutor observations.* To monitor treatment implementation fidelity, we collected data via observation forms on (a) tutor adherence to scripted *Quick Reads* protocols, (b) tutor instructional behaviors, and (c) student progress in terms of the amount of time spent actively engaged in reading passages. Tutors' fidelity to protocols was measured using a 5-point rating scale of 1 (*never*) to 5 (*always*) for each intervention step previously described. Tutor instructional behaviors were measured using the same 5-point scale for six criteria that included "maximizes time on instruction," "quick pace/smooth transitions/minimal pauses," "uses appropriate specific praise," "materials are organized," "maintains accurate



attendance records,” and “provides appropriate error correction/scaffolding.” Student progress was measured by recording the amount of time (in seconds) students were orally reading text. Prior to onsite tutor observations, we established interobserver reliability among the five researcher-observers using five videotaped *Quick Reads* sessions. Each observer viewed and rated each taped session using the observation criteria. From these ratings, interobserver reliabilities were computed. Internal consistency for the five raters was .88 for tutoring protocol fidelity and .90 for tutor instructional behaviors. Reliability for text reading time ranged from .88 to 1.00 and averaged .95 (all  $ps < .05$ ). Across 170 observations (approximately 8 per tutor and 4 per dyad), adherence to protocols averaged 4.76 ( $SD = 0.20$ ) and tutor instructional behaviors averaged 4.76 ( $SD = 0.18$ ). Each student (within their dyad) spent an estimated average of 10.4 min per session ( $SD = 3.16$ ) actively engaged in oral reading.

### Student Assessments

Students were individually assessed by trained testers unaware of group assignment on RAN and four reading subskills that were hypothesized to be differentially affected by intervention: word reading accuracy, word reading efficiency, fluency, and comprehension. Except for two submeasures of fluency, norm-referenced standard scores were used for analyses. In the measure descriptions that follow, published reliabilities are provided, as well as reliabilities for our sample (internal consistencies are Cronbach's alpha).

1. RAN was measured at pretest only using the Letter Naming subtest of the Rapid Automatized Naming/Rapid Alternating Stimulus tests (Wolf & Denckla, 2005). For this subtest, students were presented with a card that had five randomly sorted letters (*a, d, o, p, s*) repeated 10 times each and were asked to say the names of the letters as quickly as they could. The raw score was the total number of seconds the student used to name all of the letters. Test-retest reliability reported in the test manual is .87 for elementary grades. We also measured students on the Number Naming subtest to obtain concurrent validity: The correlation between Letter Naming and Number Naming (in number of seconds) was .79.

2. Word reading accuracy was measured using the Word Identification subtest from the norm-referenced, standardized Woodcock Reading Mastery Test—Revised/Normative Update, Form H (Woodcock, 1987/1998). Students were required to read increasingly difficult words, and testing was discontinued after six consecutive incorrect responses. Split-half reliability reported in the test manual averages .99 for Grades 1 to 3. Internal consistencies for our sample were .94 at pretest and .92 at posttest.

3. Word reading efficiency was measured using the Sight Word subtest from the norm-referenced, standardized Test of Word Reading Efficiency, Form B (Torgesen, Wagner, & Rashotte, 1999). The Sight Word subtest required students to read as many words as possible in 45 s from a list of increasingly difficult words. Test-retest reliability reported in the test manual for 6- to 9-year-olds is .96. For our sample, internal consistencies were .94 at pretest and .95 at posttest.

4. Fluency was the primary skill targeted in the intervention and was assessed two ways: in a raw-score framework as PRF and in a norm-referenced framework as fluency rate. Although a norm-

referenced measure of fluency can provide a generalizable measurement of students' fluency rate, we considered PRF—in terms of wcpm performance—to be a meaningful indicator of students' responsiveness to intervention. Thus, we measured students' PRF using two passages drawn from the Dynamic Indicators of Basic Early Literacy Skills: one grade-level passage used at each test interval (uniform passage; for second graders, this was “If I Had a Robot,” and for third graders, this was “My Friend”) and one grade-level alternate passage at each test interval. Alternate passages for second graders included “Roller Coaster” and “Drift Bottle”; for third graders, alternate passages included “Fieldtrip” and “Parents.” For each passage, students read aloud while the tester recorded errors, and testing was discontinued after 1 min. Words omitted, words substituted, and hesitations of more than 3 s were scored as errors. For our sample, the correlations between the uniform second- and third-grade passages were .88 and .91 at pretest and posttest, respectively, and the correlations between the uniform and alternate passages were .84 and .86 for pretest and posttest, respectively. Across passages, internal consistencies (using words as items) ranged from .96 to .98.

We assessed PRF using both uniform (PRF-U) and alternate (PRF-A) passages to minimize measurement error from potential passage nonequivalence effects (by using a uniform passage) and potential passage memory effects (by using alternate passages). Although the use of a uniform passage introduces potential memory effects (i.e., the student may remember words in a passage read on a previous occasion and thereby be able to read the text faster; or the student may remember the story structure, which may prime faster text reading) and low generalizability (i.e., the uniform passage has a specific vocabulary and story structure that differ from other passages), any memory effects occurring for treatment students have an equal chance of occurring for control students. Whereas estimates of true gains across all students on the uniform passage would likely be inflated, estimates of treatment effects on gains using the uniform passage should be free of nonequivalence effects. Consistent with our reasoning, Jenkins, Zumeta, and Dupree (2005) found that use of uniform passage measurements (compared to alternate passages) measured fluency gains more reliably; however, memory effects were detected for the uniform passage at the 5-week retest.

Our second, norm-referenced, assessment framework for measuring fluency included the Rate subtest from the Gray Oral Reading Tests—4 Form B (GORT; Wiederholt & Bryant, 2001). For this subtest, students read passages aloud, beginning at their grade level. Performance on the grade-level passage determined whether students subsequently read a lower or higher level passage, and testing was discontinued when a basal and ceiling were established. The amount of time it took students to read each passage converted to a raw score, ranging from 0 to 5 for each passage read. To be consistent with our other norm-referenced measures, we transformed the GORT standard score distribution with a mean of 10 and standard deviation of 3 to have a mean of 100 and standard deviation of 15. Cronbach's alpha reported in the test manual is .92 for 7- to 8-year-olds. Sample internal consistencies were .84 at pretest and .91 at posttest.

5. Comprehension was assessed at pretest and posttest with the GORT Comprehension subtest. Students read passages aloud, beginning at their grade level, and were asked comprehension questions for each passage. Performance on the grade-level passage



determined subsequent passage reading, and testing was discontinued when a basal and ceiling were established. Standard scores were transformed to have  $M = 100$  and  $SD = 15$ . Alternate form reliability reported in the test manual is .96 for 7- to 8-year-olds. Sample internal consistencies were .83 at pretest and .89 at post-test.

### *Classroom Literacy Instruction Observations*

We hypothesized that a specific dimension of classroom literacy instruction would interact with group assignment: time spent on OTR. For example, teachers who spend more time on OTR may bolster control students' fluency skills in the same manner as that expected for treatment students. We therefore endeavored to quantify the classroom literacy instruction for all students in the intervention study. Fifty-two classroom teachers (or reading/resource room teachers, as some students received regular reading instruction with these teachers) were asked for permission to allow us to observe their midyear reading instruction on two separate occasions (approximately 1 month apart, in one of three combinations: December/January, January/February, or February/March). Six (12%) declined to participate.

**Observation tool.** We used an adapted version of the *Instructional Content Emphasis—Revised* (ICE-R; Edmonds & Briggs, 2003) for measuring classroom literacy instruction. The standard ICE-R includes the following four dimensions: main instructional category, grouping arrangement, instructional subcategory, and materials used. We simplified the measure because establishing interobserver reliability across all dimensions would have been enormously challenging. We therefore collected data on the first two dimensions only: main instructional category and grouping arrangement.

The main instructional category measures instructional time spent on 10 mutually exclusive literacy activities: concepts of print, phonological awareness, alphabetic knowledge, word study/phonics, spelling, oral language development, fluency building, text reading, comprehension, and writing/language arts. After reviewing sample classroom reading instruction videotapes, we identified the need for four additional mutually exclusive instructional activity codes: behavior management, evaluative feedback, transition time, and other types of nonliteracy instruction (i.e., math instruction). Activities coded as behavior management included the teacher reminding students of behavior consequences, reinforcing good behavior, and talking to students who were inattentive or off task. Activities coded as evaluative feedback included explaining procedures for an instructional task and providing error correction feedback. Activities coded as transition time included students putting materials away, cleaning up, waiting, and lining up.

In addition to these 14 activities, we purposefully added a code to measure instructional time spent on OTR. Although OTR overlaps with the activities described previously (in particular, fluency building and text reading), we wished to capture time specifically afforded to OTR opportunities similar to those in *Quick Reads* instruction (in contrast to fluency building, which includes repeated word reading; and text reading, which includes silent reading practice).

The grouping arrangement dimension, which overlapped with the main instructional category activities, measured the amount of

time the teacher grouped students for instruction in whole class, small-group, pair, independent, and individualized arrangements. During each of these instructional formats, students' overall level of engagement was rated using a 3-point rating scale (1 = *low*, 2 = *moderate*, and 3 = *high engagement*).

**Recording process.** Observations were conducted for the entire duration of teachers' literacy blocks. Observers began timing at the beginning of the observation. When instruction began, observers coded the teacher's first instructional activity, and thereafter recorded each clock time (running forward from zero) associated with instructional change, along with the appropriate instructional codes. Time entries and codes were entered into a database, and the time spent on each code was computed automatically for each teacher for each observation occasion.

**Observers and reliability.** Three certificated teachers who were tutor observers also served as classroom instruction observers. Each observer studied the ICE-R manual and coding instructions and participated in several training sessions to review the measure. Observers were trained by Patricia F. Vadasy, and, to establish reliability (prior to onsite observations), observers used the ICE-R to code one 60-min videotape of language arts instruction for a first-grade classroom at one of the schools participating in the study (first grade was used because we did not wish to bias potential observations of second- and third-grade classroom teachers). To calculate reliability, we sectioned total videotape time (60 min) into 10-min intervals (the first 10 min, the second 10 min, etc.) and treated each interval as one observation. For each observer, time spent on each activity per observation was computed. Interobserver correlations were .70 or higher, and Cronbach's alpha ranged from .79 to .99.

**Observation results.** Although we had planned to conduct three observations of each classroom's literacy block, coordinating teachers' schedules with those of our small group of observers proved extraordinarily difficult. Our compromise was to conduct two observations per teacher. For each teacher, the length of time spent on each activity was averaged across the two observations. Table 2 provides a summary of teachers' average midyear time spent on each activity from the first dimension (main instructional category). Literacy blocks ranged from 32 to 132 min and averaged 76.6 min ( $SD = 21.03$ ). On average, teachers spent the majority of their time on comprehension (27% of the average amount of time), text reading, and transitioning students from one activity to another. As was expected of second- and third-grade reading instruction, teachers spent the least amount of time on concepts of print, phonological awareness, and alphabetic knowledge (i.e., less than 30 s). Eighty-five percent of average OTR time (11.9 min) overlapped with average text reading time; in other words, 66% of general text reading time was OTR. The remaining 15% of average OTR time overlapped with word study (3%), fluency building (3%), comprehension (4%), writing (3%), and evaluative feedback (3%). Teachers primarily grouped students for whole-class ( $M = 29.8$  min,  $SD = 19.35$ ) or small-group ( $M = 18.1$  min,  $SD = 18.05$ ) instruction. Teachers were also observed implementing paired (4.2 min, 5%), independent (7.6 min, 10%), and individualized (9.6 min, 13%) instructional grouping. Finally, students were observed as having high engagement 68% of the time (51.9 min) and moderate engagement 15% of the time on average. Low engagement was rare at 4% of the time.

Table 2  
Classroom Literacy Instruction Observations: Characteristics for Primary Dimension Subtypes

Name	Description	M	SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Concepts of print	Print directionality, book parts, text features	0.3	0.86	—														
2. Phonological awareness	Rhyming, syllable or phoneme blending, segmenting, isolating	0.2	0.89	-.03	—													
3. Alphabets	Letter identification and recognition	0.2	1.27	-.05	-.04	—												
4. Word study	Phonics, letter identification and recognition, decoding, sight words	6.3	7.58	-.21	-.05	-.15	—											
5. Spelling	Oral or written spelling, practice with patterns	0.9	1.87	<b>.60</b>	-.10	-.06	-.14	—										
6. Oral language	Teacher and students discuss words, books, songs, topics	1.6	2.39	.26	-.02	.17	.04	.04	—									
7. Fluency building	Students read aloud to build speed; accuracy; prosody for letters or sounds, words, text	0.8	1.73	.08	-.08	.08	.23	-.03	-.06	—								
8. Text reading	Students read orally or silently with class, small groups, individually	17.9	10.06	<b>-.31</b>	-.17	.01	.22	-.09	-.17	.19	—							
9. Comprehension	Vocabulary, prior knowledge, monitoring understanding, strategy instruction	21.0	12.64	.05	-.21	.06	-.16	.12	-.09	-.01	.13	—						
10. Writing	Composition, independent or group, mechanics, handwriting	5.0	7.18	.11	-.06	.03	.05	-.04	.09	-.01	-.26	<b>-.41</b>	—					
11. Behavior	Noninstructional time to manage behavior	1.3	2.61	-.10	-.09	.05	.02	-.11	-.07	-.08	.16	-.16	-.02	—				
12. Error correction	Teacher feedback on student errors, miscues	8.7	6.58	-.03	.12	.03	-.08	-.10	-.12	.09	-.22	<b>-.32</b>	<b>.38</b>	.01	—			
13. Other	Nonreading instruction (e.g., math concepts)	0.9	2.17	.08	-.09	-.07	-.17	.02	-.07	-.19	-.06	.05	.00	-.15	-.15	—		
14. Transitioning	Time to gain students' attention on tasks, to reorganize, or to move groups	11.5	5.06	-.03	.02	.03	.27	.17	-.17	.14	<b>.43</b>	.01	-.08	.28	-.03	-.24	—	
15. Oral text reading	Students working with teacher are reading aloud	14.0	8.79	-.22	-.21	.26	.20	-.10	-.17	<b>.33</b>	<b>.47</b>	-.08	-.18	.03	<b>.30</b>	—	-.22	.23

Note. *N* = 46 teachers; 6 teachers missing data. All categories are nonoverlapping, except oral text reading; number of minutes of literacy block reported, calculated as the average from two separate midyear observations (1 month apart). Pearson's *r* is reported; statistically significant correlations at the .05 level are in bold.



Results

Data Analysis Strategy

Due to nesting structures present in our research design, a multilevel (hierarchical) modeling approach was adopted for analyzing group differences on pretests and gains. Specifically, we tested group differences on pretests using two-level models (students within classrooms) with group dummy coded (1 = treatment and 0 = control). Two-level models were also used to analyze pretest–posttest gains; for these analyses, group was effect coded (1 = treatment and –1 = control) in order to test treatment interactions. Pretest–posttest gains were analyzed two ways: The first set of analyses tested the direct effects of the treatment on outcomes, and the second set tested the unique effects of the treatment in the presence of classroom OTR time and pretest RAN and word reading accuracy, which effectively excluded 13 students whose teachers declined observation. For the second set of models, OTR and pretest scores were grand-mean centered for ease of interpretation. In both sets of models, we ignored dyad membership because (a) assignment to dyads was cross-classroom; (b) we reasoned that dyad membership would have less impact on gains than classroom reading instruction (particularly for controls); (c) we wished to test cross-level classroom OTR effects on outcomes; and (d) although a cross-classified multilevel model can account for dyad and classroom nesting structures, cross-classified models are far more complex than hierarchical models. For all hierarchical analyses, HLM was used (Raudenbush & Bryk, 2002; Raudenbush, Bryk, Cheong, & Congdon, 2004; Rodenbush, Bryk, & Congdon, 2004); all other analyses were conducted in SPSS (1989–2004).

Pretests

Observed group means and standard deviations are reported in Table 3. According to Hasbrouck and Tindal (2006) norms, students ranged from the 10th to the 60th percentiles and averaged the 34th percentile on pretest PRF performance. According to stu-

dents’ fluency rate scores from the GORT (Wiederholt & Bryant, 2001), however, the sample averaged in the lower 10th percentile. Pretests also showed that the sample averaged in the lower 25th percentile in RAN, lower 40th percentile in word reading accuracy, lower 25th percentile in word reading efficiency, and lower 20th percentile in comprehension.

Results from our hierarchical pretest models revealed one significant difference between groups at study onset: The control group was 2.2 points higher than the treatment group on pretest word reading accuracy,  $t(160) = -2.262, p < .05$ . As expected, model results also showed significant variation between classrooms on both PRF measures and the norm-referenced fluency rate (chi-square test  $p$  values  $< .001$ ); nevertheless, no reliable between-classroom variance was detected for pretest word reading accuracy, word reading efficiency, or comprehension.

Treatment-Related Relationships

We explored the relationships between treatment students’ pretest–posttest gains and treatment-related variables, including *Quick Reads* passages read, rates of passages read per session, and whether students missed classroom reading instruction during tutoring (the majority of students were pulled out for 30 min of the typical 90-min literacy block). As shown in Appendix A, none of these variables were significantly related to student outcomes. (Although not shown in Appendix A, neither total hours of instruction nor tutor fidelity ratings were correlated with gains.) OTR, in contrast, positively affected gains on all measures except comprehension, regardless of group (see Appendix B).

Direct Treatment Effects on Pre–Post Gains

Results from our hierarchical linear models for pretest–posttest gains revealed that students’ average conditional gains were significantly greater than 0 across all measures (all  $t$ -test  $p$  values  $< .05$ ; see Table 3 for observed mean gains). Direct treatment effects were detected for gains on word reading accuracy and all three

Table 3  
Observed Pretests, Posttests, and Pretest-Posttest Gains

Measure	Treatment						Control					
	Pretest		Posttest		Gain		Pretest		Posttest		Gain	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
RAN	87.7	10.12					90.5	9.99				
WR accuracy	94.3	7.63	97.8	7.04	3.5	5.66	96.5	6.09	98.3	6.25	1.8	5.29
WR efficiency	87.9	9.19	94.7	10.12	6.8	7.72	89.2	8.50	94.6	10.49	5.5	8.09
PRF-U	40.0	18.35	83.1	22.27	43.2	16.53	42.3	19.91	79.0	23.76	36.7	19.56
PRF-A	37.5	16.84	68.3	26.38	30.8	18.50	39.6	17.13	62.5	26.63	22.9	18.36
Fluency rate	77.0	8.23	88.6	13.13	11.6	9.62	78.6	8.07	86.8	12.10	8.2	9.25
Comprehension	84.8	13.09	92.8	15.72	8.0	13.74	87.6	13.71	92.9	16.22	5.4	16.12

*Note.* Treatment group  $N = 82$ , control group  $N = 80$ . Norm-referenced standard scores (raw scores adjusted for age) were used for all measures except PRF-U and PRF-A; for these two measures, words correct per minute were used. The control group performed significantly higher on the WR accuracy pretest. RAN = Letter Naming subtest from the Rapid Automatized Naming/Rapid Alternating Stimulus tests; WR accuracy = Word Identification subtest from the Woodcock Reading Mastery Test—Revised/Normative Update; WR efficiency = Sight Word subtest from the Test of Word Reading Efficiency; PRF-U and PRF-A = words-correct-per-minute performance on uniform and alternate passages, respectively, from the Oral Reading Fluency subtest from the Dynamic Indicators of Basic Early Literacy Skills; Fluency rate = Rate subtest from the Gray Oral Reading Tests-4 (GORT); Comprehension = GORT Comprehension subtest.

fluency measures but not for word reading efficiency or comprehension: On average, students in the treatment group gained 1.6 more standard score points than controls on word reading accuracy,  $t(160) = 2.206, p < .05$ ; and 2.6 more points on fluency rate,  $t(160) = 2.537, p < .05$ . On PRF-U, treatment students had an estimated 5.8-wcpm advantage over controls,  $t(160) = 2.276, p < .05$ ; and a 7.9-wcpm advantage on PRF-A,  $t(160) = 3.008, p < .01$ . Model results also revealed significant between-classroom variation on gains for all measures except comprehension (chi-square test  $p$  values  $< .05$ ). Taking the square root of the total estimated variance (between-classroom + residual variance) as a pooled standard deviation, we calculated approximate treatment effect sizes (Cohen's  $d$ ) and found  $d = .29, .31, .32$ , and  $.43$ , for gains in word reading accuracy, fluency rate, PRF-U, and PRF-A, respectively.

### *Conditional Treatment Effects on Pre-Post Gains*

Our second, more complex set of hierarchical models tested the unique effect of treatment on student gains in the presence of classroom OTR time, pretest RAN, pretest word reading accuracy (WR), and corresponding interactions. As a reminder, these analyses excluded 13 cases due to the six teachers who declined to participate in classroom observations. As a check on the potential bias introduced into our models due to excluding these cases, we compared students who were excluded with those who were included and found no differences between groups on any variable included in our models: All  $t$ -test  $p$  values  $> .05$  for RAN, WR, and all gains. As shown in Table 4, results from these models revealed significant unique treatment effects on word reading efficiency (approximate  $d = .32$ ), PRF-U ( $d = .65$ ), PRF-A ( $d = .81$ ), fluency rate ( $d = .58$ ), and comprehension ( $d = .51$ ). The treatment effect on word reading accuracy gains detected in our direct effects model was now accounted for by the other variables (recall that the treatment group was significantly lower than the control group on WR pretest). Moreover, although direct treatment effects on word reading efficiency and comprehension were not detected in our prior direct effects models, unique positive treatment effects were found in the presence of OTR, RAN, WR, and interactions.

**Unique OTR effects.** Model results revealed that OTR had small but reliable unique effects on students' gains in word reading accuracy ( $d = .30$ ), word reading efficiency ( $d = .33$ ), PRF-U ( $d = .33$ ), and fluency rate ( $d = .36$ ). (To compute the effect sizes, we contrasted model-predicted student gains for 1  $SD$  above the mean, OTR = 1, with student gains at the average, OTR = 0.) For example, students whose teachers spent 1  $SD$  more on OTR (approximately 22.8 min; see Table 2) were estimated to have a 1.3-point advantage on word reading accuracy gain over those whose teachers spent less time (5.2 min), holding other variables constant.

**Unique pretest effects.** Results showed that RAN was uniquely predictive of higher gains in word reading accuracy ( $d = .20$ ), both PRF measures ( $d = .38$  and  $.39$  for PRF-U and PRF-A, respectively), and fluency rate ( $d = .26$ ). WR was also uniquely predictive: Students with higher WR were predicted to have smaller word reading accuracy gains ( $d = -.49$ ) but higher gains in word reading efficiency ( $d = .23$ ), PRF-U ( $d = .27$ ), and fluency rate ( $d = .36$ ).

**Interactions.** There were complex interactions present across all gain measures except PRF-A. Although there were no signif-

icant two-way Group  $\times$  OTR interactions, there were three significant complex interactions involving group, OTR, and pretest measures (for PRF-U, there was a significant Group  $\times$  RAN  $\times$  OTR interaction; for comprehension, there was a significant Group  $\times$  WR  $\times$  OTR interaction; for both, there were significant four-way interactions). Significant interactions between group and RAN (for gains on PRF-U and comprehension) and between group and WR (for gains on word reading accuracy and efficiency, PRF-U, and fluency rate) were evident, as were interactions between RAN and OTR (for gains on word reading accuracy, both PRF measures, and comprehension) and RAN and WR (for gains on word reading accuracy and efficiency, and comprehension). In general, these interactions indicate that the benefits of a repeated reading intervention and/or increased classroom OTR for second- and third-grade students is dependent on students' initial RAN and WR skills.

To achieve better insight about these interactions, we created plots (see Figures 1–3) using model-predicted parameter estimates (see Table 4) in concert with selected values of the predictors as follows: Group was 1 = treatment,  $-1$  = control; OTR was  $-1$  = 1  $SD$  below the mean (5.2 min of the literacy block), 0 = mean (14.0 min of the literacy block), and 1 = 1  $SD$  above the mean (22.8 min of the literacy block); pretest RAN was  $-1$  = 1  $SD$  below the mean (10th percentile based on norms), 0 = mean (25th percentile based on norms), and 1 = 1  $SD$  above the mean (50th percentile based on norms); and pretest WR (all values observed in the sample, from 76 to 114). Visual inspection revealed a few remarkable patterns. First, OTR practice—whether in the classroom (OTR) or in paired tutoring (treatment)—appeared to benefit students who were below (approximately) the 50th percentile on fall WR (irrespective of pretest RAN) on gains in word reading accuracy and word reading efficiency. This pattern was similar for gains on both PRF-U and fluency rate measures, with one exception. Students with very low RAN (bottom 10th percentile) and low WR appeared to have higher gains if they received both treatment and high OTR or received no treatment with low OTR (perhaps more classroom time allocated to word study development). Finally, inspection of Figure 3 (right side) illustrates the complexity of individual differences in comprehension gains: Students with very low to moderately low RAN (10th to 25th percentile) appeared, overall, to benefit from paired tutoring over no tutoring; that said, treatment students did not appear to benefit from added OTR in terms of comprehension gains, whereas control students did appear to benefit from added OTR. Taken together, these findings suggest that students with lower RAN will have the highest comprehension gains if only a moderate amount of OTR practice is employed. Nevertheless, students with average RAN did not appear to be differentiated by group in terms of comprehension; instead, it appeared that less OTR had positive benefits for these students.

### Discussion

The present study evaluated the direct and indirect effects of a supplemental, paraeducator-implemented repeated reading intervention (*Quick Reads*) with incidental word-level instruction for second and third graders with low fluency skill. Results show clearly that students benefited from this intervention in terms of word reading and fluency gains. Specifically, our models that



Table 4  
Conditional Effects of Group, Classroom Oral Text Reading, and Pretests on Student Gains

Effect	WR accuracy			WR efficiency			PRF-U			PRF-A			Fluency rate			Comprehension		
	Coefficient	SE	<i>t</i> (125)	Coefficient	SE	<i>t</i> (125)	Coefficient	SE	<i>t</i> (125)	Coefficient	SE	<i>t</i> (125)	Coefficient	SE	<i>t</i> (125)	Coefficient	SE	<i>t</i> (125)
Fixed effects																		
Conditional mean gain	2.70	0.47	5.8***	5.97	0.75	7.9***	38.83	1.38	28.1***	27.09	1.67	16.2***	9.45	0.75	12.6***	6.67	1.17	5.7***
Group <sup>a</sup>	0.63	0.41	1.6	1.11	0.46	2.4*	4.93	1.43	3.5***	6.53	1.02	6.4***	2.37	0.65	3.6***	3.24	1.17	2.8**
OTR	1.33	0.39	3.5**	2.25	0.69	3.2**	5.01	1.26	4.0***	2.71	1.50	1.8	2.94	0.63	4.6***	-0.22	1.00	-0.2
RAN	0.89	0.39	2.3*	0.77	0.63	1.2	5.77	1.07	5.4***	6.26	1.73	3.6***	2.10	0.92	2.3*	1.15	1.12	1.0
WR	-2.23	0.41	-5.5***	1.59	0.60	2.6*	5.66	1.97	2.9*	2.21	1.95	1.1	2.95	0.94	3.1**	-0.74	1.28	-0.6
Group × OTR	0.35	0.34	1.0	0.24	0.44	0.5	-0.94	1.29	-0.7	-0.65	0.81	-0.8	0.36	0.61	0.6	-1.91	1.22	-1.6
Group × RAN	-0.51	0.38	-1.4	-0.22	0.51	-0.4	-3.57	1.30	-2.8**	-1.26	1.20	-1.0	0.19	0.87	0.2	-3.86	1.32	-2.9**
Group × WR	-0.83	0.42	-2.0*	-1.98	0.55	-3.6**	-3.76	1.72	-2.2*	-3.17	2.08	-1.5	-1.89	0.88	-2.1*	-1.91	1.63	-1.2
OTR × RAN	0.75	0.27	2.7**	0.49	0.52	0.9	2.56	0.73	3.5***	3.03	1.24	2.4*	1.35	0.72	1.9	3.76	1.10	3.4***
OTR × WR	0.33	0.34	1.0	0.81	0.55	1.5	1.01	1.79	0.6	0.68	1.95	0.3	0.57	0.96	0.6	-0.97	1.44	-0.7
RAN × WR	-1.16	0.54	-2.1*	-1.54	0.65	-2.4*	-2.24	1.69	-1.3	-2.23	1.71	-1.3	-1.59	0.84	-1.9	-4.32	1.59	-2.7**
Group × OTR × RAN	-0.48	0.33	-1.4	-0.50	0.43	-1.2	-0.85	0.68	-1.3	-0.82	1.16	-0.7	-0.01	0.70	0.0	-3.40	1.45	-2.3*
Group × OTR × WR	-0.65	0.36	-1.8	-0.84	0.48	-1.7	-3.56	1.52	-2.3*	-3.66	2.14	-1.7	-1.46	0.81	-1.8	-0.46	1.48	-0.3
Group × RAN × WR	-0.07	0.50	-0.2	0.85	0.62	1.4	1.08	1.93	0.6	-1.42	1.88	-0.8	0.42	1.18	0.4	1.71	1.43	1.2
OTR × RAN × WR	-0.76	0.42	-1.8	-0.85	0.63	-1.4	-3.75	1.16	-3.2**	-1.73	1.16	-1.5	-1.26	0.71	-1.8	-3.69	1.53	-2.4*
Group × OTR × RAN × WR	0.41	0.35	1.2	0.48	0.60	0.8	3.19	1.11	2.9**	0.64	1.46	0.4	0.39	0.83	0.5	5.24	1.27	4.1***
Random effects	Var	SD	$\chi^2(44)$	Var	SD	$\chi^2(44)$	Var	SD	$\chi^2(44)$	Var	SD	$\chi^2(44)$	Var	SD	$\chi^2(44)$	Var	SD	$\chi^2(44)$
Classrooms, <i>U</i> <sub>0</sub>	0.7	0.86	55.7	9.6	3.10	83.7***	24.1	4.91	67.0*	49.0	7.00	78.6***	4.2	2.06	56.8	0.1	0.38	37.4
Residual, <i>e</i>	19.5	4.42		37.5	6.12		204.8	14.31		214.0	14.63		63.3	7.96		162.3	12.74	

Note. *N* = 46 teachers and 141 students (*n* = 71 treatment, *n* = 70 control). Degrees of freedom for conditional mean gains is 44; *df* = 125 otherwise. OTR = standardized (grand-mean centered) average midyear minutes spent on oral text reading; RAN = standardized (grand-mean centered) pretest Letter Naming standard score from the Rapid Automatized Naming/Rapid Alternating Stimulus tests. WR accuracy (as predictor) = standardized (grand-mean centered) pretest Word Identification standard score from the Woodcock Reading Mastery Test—Revised/Normative Update (WRMT-R/NU). For each outcome: pretest-posttest gains are analyzed; WR accuracy = WRMT-R/NU Word Identification; WR efficiency = Sight Word subtest from the Test of Word Reading Efficiency; PRF-U and PRF-A = words-correct-per-minute performance on uniform and alternate passages, respectively, from the Oral Reading Fluency subtest from the Dynamic Indicators of Basic Early Literacy Skills; Fluency rate = Rate subtest from the Gray Oral Reading Tests-4 (GORT); Comprehension = GORT Comprehension subtest.

<sup>a</sup> 1 = treatment; -1 = control.

\* *p* < .05. \*\* *p* < .01. \*\*\* *p* < .001.

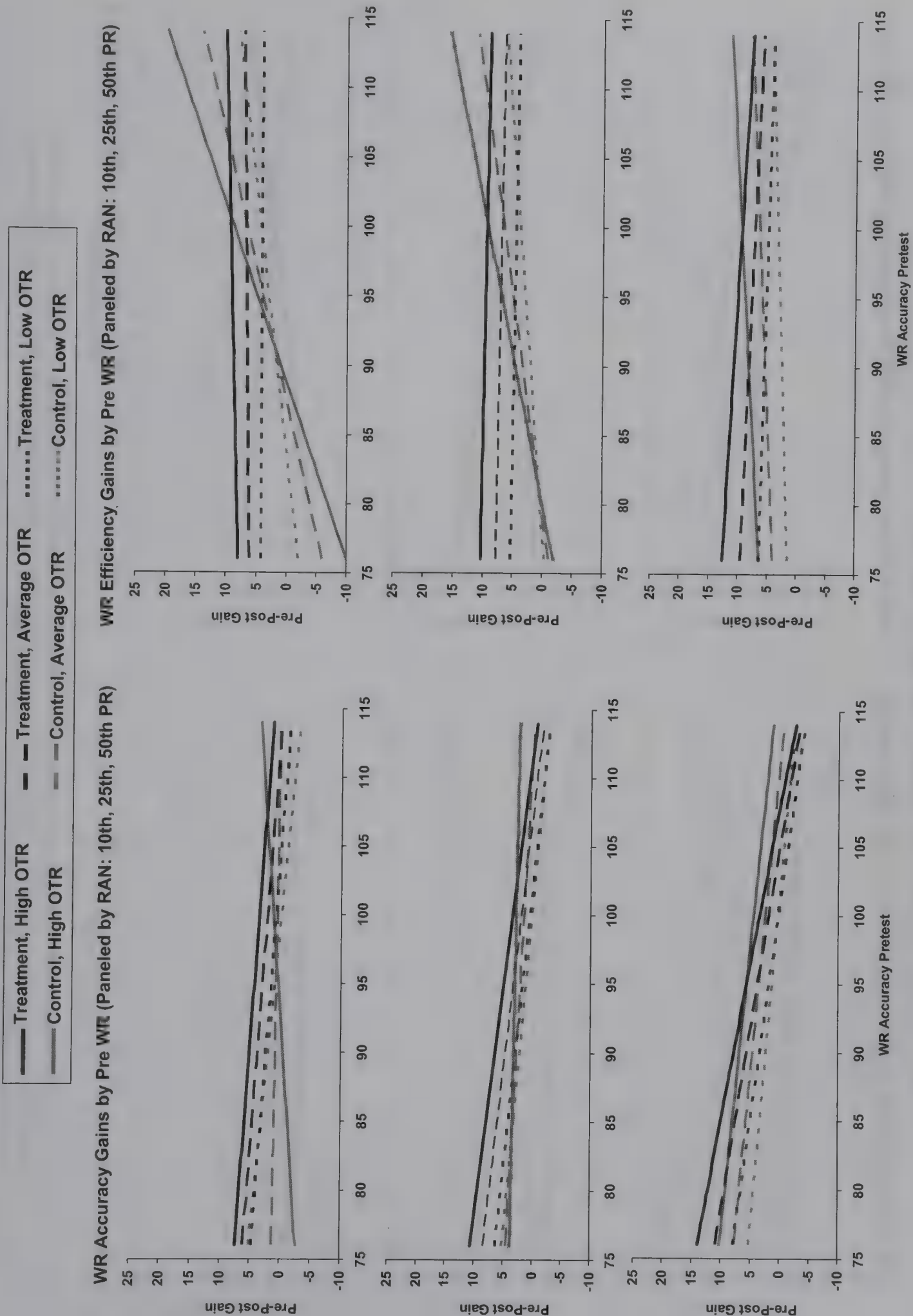


Figure 1. Model-predicted word reading (WR) accuracy and efficiency gains by pretest WR accuracy, paneled by three levels of rapid automatized naming (RAN): 10th percentile (PR; top), 25th PR (middle), and 50th PR (bottom). Lines represent experimental group and amount of classroom oral text reading (OTR).





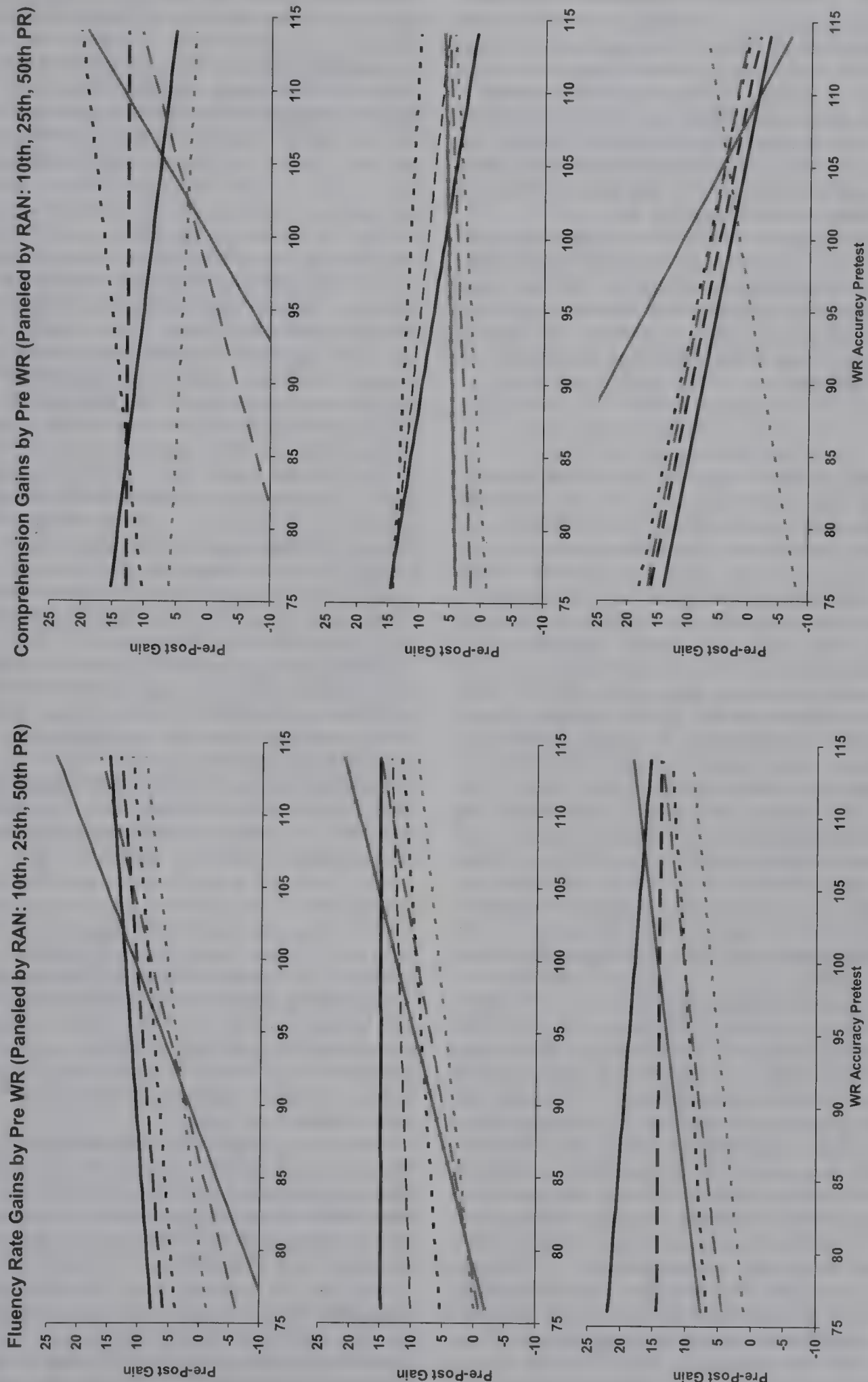
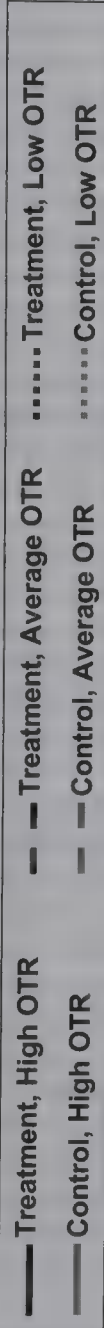


Figure 3. Model-predicted fluency rate and comprehension gains by pretest word reading (WR) accuracy, paneled by three levels of rapid automatized naming (RAN): 10th percentile (PR; top), 25th PR (middle), and 50th PR (bottom). Lines represent experimental group and amount of classroom oral text reading (OTR).



tested for direct treatment effects indicated that tutored students had significantly higher pretest–posttest gains in word reading accuracy and fluency. Dividing the treatment group's average wcpm gains (see Table 3) by 15 weeks of intervention yields average weekly gains of 2.9 and 2.1 wcpm for PRF–U and PRF–A, respectively. By both counts, this is well over the 1.1 wcpm/week expected for low to moderately low students in Hasbrouck and Tindal (2006). Nevertheless, the treatment group's mean posttest PRF performance was below the 50th percentile according to all three of our fluency measures (see Table 3).

Of our classroom observations, the characteristic of most interest in this study was the amount of time students spent reading aloud during the literacy block. When classroom OTR time, pretest RAN, and pretest WR were added to our second set of models, significant unique treatment effects were detected for gains on word reading efficiency, fluency, and comprehension, but not word reading accuracy.

### Gains

*Word reading accuracy and efficiency gains.* Children entered this study with fluency performance ranging from the 10th to 60th percentiles according to Hasbrouck and Tindal (2006) norms on two grade-level passages, and in the bottom 10th percentile according to norms on the Rate subtest from the GORT. Verbal efficiency theory (Perfetti, 1992; Perfetti & McCutchen, 1987) suggests that students' beginning word reading skill and naming speed influence the consolidation of specific fluency component skills. The significant, complex interactions observed in our second set of models unmistakably suggest that effects of repeated reading practice on reading subskill gains are dependent on beginning word reading and/or RAN skills. Examination of the findings for word reading accuracy gains (see Table 4) showed that the repeated reading intervention with incidental word-level instruction (treatment) was reliably moderated by pretest WR: Holding the other variables constant, treatment students with lower pretest WR were predicted to have a slightly higher gain in word reading accuracy compared with controls, whereas the opposite was predicted to be true for students with higher pretest WR. Classroom OTR, pretest RAN, and their interaction had unique, positive effects on word reading accuracy gains. This suggests that students with higher RAN who receive more classroom OTR will have a cumulative advantage of approximately 3 points ( $1.33 + 0.89 + 0.75 = 2.97$ ) in gains compared with students with lower RAN who receive lower amounts of classroom OTR, holding all other variables constant. Finally, this outcome was the only one in which pretest WR had a unique negative effect on gains, indicating that students with higher WR pretest had lower word reading accuracy growth. Furthermore, pretest WR interacted negatively with RAN in addition to group, revealing that students with low pretest WR and high RAN had higher gains. Findings for word reading efficiency gains (see Table 4 and Figure 1, right side) were similar to those found for word reading accuracy, although for this outcome treatment had a unique, positive effect and there were no unique RAN effects or interactions between RAN and classroom OTR.

*Fluency gains.* Reading fluency rate gains were measured in a raw-score and norm-referenced framework using grade-level reading passages (one uniform passage, PRF–U; and one alternate passage, PRF–A) and the norm-referenced GORT Rate subtest.

For all three measures, treatment students had significantly higher gains than controls, and, similarly, students with higher RAN had higher gains. Except for PRF–A, classroom OTR and pretest WR were positively predictive of fluency gains; however, just like our models for word reading efficiency and accuracy, a significant negative interaction between pretest WR and experimental group was also detected: Treatment students with lower pretest WR were predicted to have higher gains than treatment students with higher pretest WR. Two other findings with respect to fluency gains are noteworthy. First, for both of our PRF passages (PRF–U and PRF–A), the benefits of effects of classroom OTR were positively dependent on pretest RAN: In other words, students with higher RAN had more raw score fluency gains if they were in classrooms with higher amounts of OTR. This interaction is consistent with our findings for word reading accuracy. Second, gains on our uniform passage fluency measure (PRF–U) resulted from reliable complex interactions between all predictors (see Table 4); although several interpretations of the three- and four-way interactions may be used, the predicted values of PRF–U gains illustrated in Figure 2 are remarkably similar to the other fluency measures.

*Comprehension gains.* After controlling for students' pretest levels and oral classroom text reading time, we detected a positive unique effect of treatment on comprehension gains. Nevertheless, this benefit of repeated reading intervention is qualified by several complex interactions among the other predictors in the model, highly similar to those detected for the PRF–U fluency measure. Predicted gain values (illustrated Figure 3, right side) appear to suggest that students with lower levels of RAN (10th or 25th percentiles) had higher gains if they were in the treatment group, and even higher gains, on average, if they were in classrooms that focused less time on OTR. In contrast, students with higher RAN (50th percentile) did not appear well-differentiated by which experimental group they were in.

Verbal efficiency theory (Perfetti, 1992; Perfetti & McCutchen, 1987) incorporates the notion of the autonomous lexicon, a bank of accurate word representations that underlie fluent reading. Findings from this study suggest that if students have lower word reading accuracy skills and an insufficiently encapsulated lexicon, higher RAN permits them to benefit from either type of oral reading practice across word and rate measures. Children with lower RAN and lower word reading accuracy had less fluency or comprehension gains from classroom OTR opportunities alone but were most advantaged by repeated reading instruction. Students' initial higher word reading skill uniquely predicted lower word reading accuracy gains but greater fluency rate gains. The role of word reading skill in explaining growth for students with low levels of reading fluency is consistent with previous studies (Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003; Torgesen et al., 2001).

These findings suggest when particular types of repeated reading will be beneficial. For students performing within the emergent stage of reading fluency, like the second and third graders in this study, there were benefits from rereading the same texts chosen at an instructional level of difficulty. Repeated reading was assisted by trained tutors who provided both a model of fluent reading during two of the rereadings, as well as corrections and feedback during three of the readings. The intervention incorporated features found by others to be effective. These include use of a structured, research-based protocol; inclusion of alphabetic and word-level instruction; and rereading of texts. Although the supplemental



intervention in the present study was not explicitly coordinated with classroom instruction, as others have recommended (Jenkins & Jenkins, 1987; Reisner et al., 1990; Venezky & Jain, 1996; Wasik, 1998a), *Quick Reads* texts are designed to coordinate with grade-level science and social science content and vocabulary. Furthermore, the two-level models in this study suggest that added oral reading practice in the classroom enhances fluency effects of the repeated reading intervention.

### Limitations

Findings from this study should be considered in light of several limitations. First, although the intervention used in this study was primarily characterized as repeated reading, a small portion (up to 5 min) of each of the tutoring sessions included incidental alphabetic instruction and word-level scaffolding. Although most repeated reading interventions described in the literature do not include this feature, the teacher manual suggests that this is an option teachers might use in implementing *Quick Reads*. Second, students entered this study with a wide range of pretest fluency levels that reflected teacher referral patterns; nevertheless, students ranged from the 10th to the 60th percentiles on PRF performance, similar to students served in repeated reading programs (Faulkner & Levy, 1999; Hasbrouck et al., 1999). Third, we observed classroom instruction only twice during the intervention. As others have demonstrated (Connor, Morrison, & Petrella, 2004), dimensions of classroom instruction that our coding system did not capture, such as individual student engagement or quality of instruction (Connor, Son, Hindman, & Morrison, 2005), may have influenced student outcomes. Fourth, our findings on classroom literacy instruction are based on data excluding six teachers (and their students). It is possible that these teachers' refusal to participate reflects a systematic difference in their literacy instruction; however, outcomes of students within these classrooms did not reliably differ from outcomes of students whose teachers were observed. A final limitation of this study is that many variables expected to contribute to comprehension gains were not accounted for in this study, including vocabulary knowledge, strategy skills, and general language skills.

### Instructional Implications

The fluency gains on both the norm-referenced and the PRF-U measure resulting from the supplemental repeated reading intervention were greatest for students with below-grade-level word reading skills. Higher levels of RAN appeared to allow students at this emergent stage of fluency development (Schwanenflugel et al., 2006) to benefit more from treatment, although this type of interaction was only reliably evident for one of our fluency measures (PRF-U). Inspection of the plots in Figures 1–3 suggests that if students enter second or third grade with average or higher word reading skills, a repeated reading intervention is not recommended to boost word reading accuracy or comprehension skills in particular. Students with lower word reading skills at onset, however, will likely benefit from this type of intervention on reading accuracy and fluency rate gains.

Second and third graders entered this study at different levels of component skill fluency. Word reading skills at pretest averaged in the lowest 40th percentile. The pretest range in word reading skills suggested that we add an incidental word-level instruction layer to the

intervention; this instruction, together with the *Quick Reads* text features, may have enabled students with lower word reading skills to make the greatest gains relative to controls across reading outcomes. Gains in fluency and comprehension reflect, to some extent, an interactive compensatory mechanism (Rumelhart, 1977) for RAN and word reading skills. It appears that when bottom-up processing was slow due to both underlying slow RAN and inaccurate decoding, students could not make gains in fluency or comprehension from classroom oral reading practice opportunities. However, either high RAN or stronger word reading skill seems to enable compensatory processing and gains in fluency and comprehension. Students with high word reading pretest made high fluency and comprehension gains despite low RAN, and students with high RAN made comprehension gains despite low word reading. Higher RAN was associated with similar comprehension gains for both treatment and control students with low word reading pretest scores.

Findings support clear benefits from the opportunities students had to engage in oral reading practice during the classroom reading block. When students read aloud, teachers have opportunities to detect student difficulties, including poor prosody, decoding errors, and limited comprehension reflected in dysfluent reading. Teachers can use this information to adjust instruction for individual students and provide effective corrections and scaffolding. Teachers have no access to this information when students engage in silent reading.

The findings also support the benefits of supplemental repeated reading instruction using the *Quick Reads* program. Students with below-grade-level word reading skill made the greatest gains in reading skills due to intervention. As we noted, tutors implemented *Quick Reads* instruction with incidental word-level instruction. We believe that this adaptation is in line with instructional applications recommended in the *Quick Reads* teacher manual and is most appropriate for students who have not yet developed accurate and efficient word reading skills. These students are also most likely to benefit from the text characteristics of the *Quick Reads* program. Students with average or higher word reading accuracy may have benefited most from incidental and more individualized instruction in fluency and comprehension skills during classroom reading instruction.

Tutors in this study were paraeducators, and they implemented the intervention with a high degree of fidelity. There is growing evidence that teaching assistants and tutors can assume important pedagogical roles in early literacy instruction (Al Otaiba, Schatschneider, & Silverman, 2005; Gelzheiser, 2005; Hatcher et al., 2006), with corresponding calls for increased training and utilization of these individuals (Callas, 2001; Savage & Carless, 2005). Research is warranted on the specific instruction and curricula these individuals can implement most effectively.

This study specifically addressed limitations in previous research on repeated reading interventions. First, students were randomly assigned to conditions. Second, we considered multiple outcomes, including word reading accuracy and efficiency as well as fluency rate and comprehension outcomes. Third, because this was an efficacy trial, the intervention was implemented with a high degree of fidelity by paraeducators who were potential typical end users, and in school settings that reflected routine practice conditions. Fourth, the 15-week intervention was considerably more intense than the repeated reading interventions described in many previous studies. Fifth, the particular repeated reading intervention, *Quick Reads*, is unusually well specified in terms of text features and



reading procedures often hypothesized to influence fluency outcomes. Finally, we conducted classroom observations to measure and account for the influence of classroom literacy instruction similar to that of the intervention on student outcomes. Specifically, we estimated OTR practice that students received in the classroom.

In summary, findings indicate that student word reading and RAN skills should be taken into account when identifying students for placement in a repeated reading intervention. The findings also underscore the value of classroom oral reading practice for students with low to moderate reading fluency and the importance of accounting for the role of classroom reading instruction experienced by children in supplemental reading interventions. Finally, findings support the difficulty described by others (Torgesen et al., 2001) of completely or quickly remediating poor fluency. This difficulty is not limited only to older students with cumulative fluency lags, as in this study we were unable to bring second- and third-grade students in the intervention up to grade-level fluency. Young students whose fluency is already constrained by limited word-level skills may require significantly greater amounts of repeated reading practice than employed in this intervention, as well as more intensive instruction to develop word-level proficiency.

## References

- Al Otaiba, S., Schatschneider, C., & Silverman, E. (2005). Tutor-assisted intensive learning strategies in kindergarten: How much is enough? *Exceptionality*, 13, 195–208.
- Berninger, V. W., Abbott, R. D., Billingsley, F., & Nagy, W. (2001). Processes underlying timing and fluency of reading: Efficiency, automaticity, coordination, and morphological awareness. In M. Wolf (Ed.), *Dyslexia, fluency, and the brain* (pp. 383–414). Timonium, MD: York Press.
- Blachman, B. A., Ball, E. W., Black, R., & Tangel, D. M. (1994). Kindergarten teachers develop phoneme awareness in low-income, inner-city classrooms: Does it make a difference? *Reading and Writing*, 6, 1–17.
- Blachman, B. A., Tangel, D. M., Ball, E. W., Black, R., & McGraw, C. K. (1999). Developing phonological awareness and word recognition skills: A two-year intervention with low-income, inner-city children. *Reading and Writing*, 11, 239–273.
- Callas, M. (2001). Current and proposed special education legislation. *Child Psychology and Psychiatry Review*, 6, 24–30.
- Cameron, C. E., Connor, C. M., & Morrison, F. J. (2005). Effects of variation in teacher organization on classroom functioning. *Journal of School Psychology*, 43, 61–85.
- Chard, D. J., Vaughn, S., & Tyler, B. (2002). A synthesis of research on effective interventions for building reading fluency with elementary students with learning disabilities. *Journal of Learning Disabilities*, 35, 386–406.
- Connor, C. M., Morrison, F. J., & Katch, L. E. (2004). Beyond the reading wars: Exploring the effect of child-instruction interactions on growth in early reading. *Scientific Studies of Reading*, 8, 305–336.
- Connor, C. M., Morrison, F. J., & Petrella, J. N. (2004). Effective reading comprehension instruction: Examining Child  $\times$  Instruction interactions. *Journal of Educational Psychology*, 96, 682–698.
- Connor, C. M., Son, S.-H., Hindman, A. H., & Morrison, F. J. (2005). Teacher qualifications, classroom practices, family characteristics, and preschool experience: Complex effects on first graders' vocabulary and early reading outcomes. *Journal of School Psychology*, 43, 343–375.
- Edmonds, M., & Briggs, K. L. (2003). The Instructional Content Emphasis Instrument: Observations of reading instruction. In S. Vaughn & K. L. Briggs (Eds.), *Reading in the classroom: Systems for the observation of teaching and learning* (pp. 31–52). Baltimore: Brookes.
- Faulkner, H. J., & Levy, B. A. (1999). How text difficulty and reader skill interact to produce differential reliance on word and context overlap in reading transfer. *Journal of Experimental Child Psychology*, 58, 1–24.
- Fleisher, L. S., Jenkins, J. R., & Pany, D. (1979). Effects on poor readers' comprehension of training in rapid decoding. *Reading Research Quarterly*, 15, 30–48.
- Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology*, 90, 37–55.
- Fuchs, D., Fuchs, L. S., Thompson, A., Al Otaiba, S., Yen, L., Yang, N. J., et al. (2001). Is reading important in reading-readiness programs? A randomized field trial with teachers as program implementers. *Journal of Educational Psychology*, 93, 251–267.
- Fuchs, D., Fuchs, L. S., Thompson, A., Al Otaiba, S., Yen, L., Yang, N. J., et al. (2002). Exploring the importance of reading programs for kindergarten children with disabilities in mainstream classrooms. *Exceptional Children*, 68, 295–311.
- Gelzheiser, L. M. (2005). Maximizing student progress in one-to-one programs: Contributions of texts, volunteer experience, and student characteristics. *Exceptionality*, 13, 229–243.
- Good, R. H., & Kaminski, R. A. (2002). *Dynamic Indicators of Basic Early Literacy Skills (DIBELS)* (6th ed). Eugene, OR: Institute for the Development of Educational Achievement.
- Gunn, B., Biglan, A., Smolkowski, K., & Ary, D. (2000). The efficacy of supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school. *Journal of Special Education*, 34, 90–103.
- Gunn, B., Smolkowski, K., Biglan, A., & Black, C. (2002). Supplemental instruction in decoding skills for Hispanic and non-Hispanic students in early elementary school: A follow-up. *Journal of Special Education*, 36, 69–79.
- Gunn, B., Smolkowski, K., Biglan, A., Black, C., & Blair, J. (2005). Fostering the development of reading skill through supplemental instruction: Results for Hispanic and non-Hispanic students. *Journal of Special Education*, 39, 66–85.
- Hasbrouck, J. E., Ihnot, C., & Rogers, G. H. (1999). "Read Naturally": A strategy to increase oral reading fluency. *Reading Research and Instruction*, 39, 27–38.
- Hasbrouck, J., & Tindal, G. A. (2006). ORF norms: A valuable assessment tool for reading teachers. *Reading Teacher*, 59, 636–644.
- Hatcher, P. J., Goetz, K., Snowling, M. J., Hulme, C., Gibbs, S., & Smith, G. (2006). Evidence for the effectiveness of the Early Literacy Support programme. *British Journal of Educational Psychology*, 76, 351–367.
- Hiebert, E. H. (2003). *Quick reads*. Parsippany, NJ: Pearson Learning.
- Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, 95, 719–729.
- Jenkins, J. R., & Jenkins, L. M. (1987, March). Making peer tutoring work. *Educational Leadership*, 44(6), 64–68.
- Jenkins, J. R., Zumeta, R., & Dupree, O. (2005). Measuring gains in reading ability with passage reading fluency. *Learning Disabilities Research and Practice*, 20, 245–253.
- Juel, C. (1996). What makes literacy tutoring effective? *Reading Research Quarterly*, 31, 268–289.
- Juel, C., & Roper-Schneider, D. (1985). The influence of basal readers on first grade reading. *Reading Research Quarterly*, 20, 134–152.
- Kame'enui, E. J., Simmons, D. C., Good, R. H., & Harn, B. A. (2001). The use of fluency-based measures of early identification and evaluation of intervention efficacy in schools. In M. Wolf (Ed.), *Dyslexia, fluency, and the brain* (pp. 307–332). Timonium, MD: York Press.
- Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology*, 95, 3–21.
- LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6, 293–323.
- Levy, B. A., Abello, B., & Lysynchuk, L. (1997). Transfer from word training to reading in context: Gains in reading fluency and comprehension. *Learning Disability Quarterly*, 20, 173–188.
- Menon, S., & Hiebert, E. H. (2005). A comparison of first graders' reading

with little books or literature-based basal anthologies. *Reading Research Quarterly*, 40, 12–38.

- Mercer, C. D., Campbell, K. U., Miller, M. D., Mercer, K. D., & Lane, H. B. (2000). Effects of a reading fluency intervention for middle schoolers with specific learning disabilities. *Learning Disabilities Research and Practice*, 15, 179–189.
- Meyer, M. S., & Felton, R. H. (1999). Repeated reading to enhance fluency: Old approaches and new directions. *Annals of Dyslexia*, 49, 283–306.
- National Reading Panel. (2000). *Report of the subgroups: National Reading Panel*. Washington, DC: National Institute of Child Health and Development.
- No Child Left Behind Act of 2001, 20 U.S.C. 70 § 6301 *et seq.* (2002).
- O'Shea, L. J., Sindelar, P. T., & O'Shea, D. J. (1985). The effects of repeated readings and attentional clues on reading fluency and comprehension. *Journal of Reading Behavior*, 17, 129–142.
- Perfetti, C. A. (1985). *Reading ability*. New York: Oxford University.
- Perfetti, C. A. (1992). The representation problem in reading acquisition. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 145–174). Hillsdale, NJ: Erlbaum.
- Perfetti, C. A., & McCutchen, D. (1987). Schooled language competence: Linguistic abilities in reading and writing. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics* (Vol. 2, pp. 105–141). Cambridge, England: Cambridge University Press.
- Rashotte, C. A., MacPhee, K., & Torgesen, J. K. (2001). The effectiveness of a group reading instruction programs with poor readers in multiple grades. *Learning Disability Quarterly*, 24, 119–134.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.
- Raudenbush, S. W., Bryk, A. S., & Congdon, R. T. (2004). *HLM for Windows 6.0* [Computer software]. Lincolnwood, IL: Scientific Software International.
- Reisner, E. R., Petry, C. A., & Armitage, M. (1990). *A review of programs involving college students as tutors or mentors in grades K-12*. Washington, DC: U.S. Department of Education.
- Roth, S. F., & Beck, I. L. (1987). Theoretical and instructional implications of the assessment of two microcomputer word recognition programs. *Reading Research Quarterly*, 22, 197–218.
- Rumelhart, D. E. (1977). Toward an interactive model of reading. In S. Dornic (Ed.), *Attention and performance* (Vol. 6, pp. 573–603). Hillsdale, NJ: Erlbaum.
- Samuels, S. J. (1979). The method of repeated readings. *Reading Teacher*, 32, 403–408.
- Samuels, S. J., Dahl, P., & Archwamety, T. (1974). Effect of hypothesis/test training on reading skill. *Journal of Educational Psychology*, 66, 835–844.
- Savage, R., & Carless, S. (2005). Learning support assistants can deliver effective reading interventions for 'at-risk' children. *Educational Research*, 47, 45–61.
- Schwanenflugel, P. J., Meisinger, E. B., Wisenbaker, J. M., Kuhn, M. R., Strauss, G. P., & Morris, R. D. (2006). Becoming a fluent and automatic reader in the early elementary years. *Reading Research Quarterly*, 41, 496–522.
- Simmons, D. C., Kame'enui, E. J., Stoolmiller, M., Coyne, M., & Harn, B. (2003). Accelerating growth and maintaining proficiency: A two-year intervention study of kindergarten and first-grade children at risk for reading difficulties. In B. R. Foorman (Ed.), *Preventing and remediating reading difficulties: Bringing science to scale* (pp. 197–228). Timonium, MD: York Press.
- Smith, V., & Siegel, L. S. (2006). *Early literacy instruction in kindergarten: Examining the implementation of a phonological awareness program in routine practice*. Unpublished manuscript, University of Alberta, Edmonton, Alberta, Canada.
- SPSS. (1989–2004). *Statistical Package for the Social Sciences 13.0* [Computer software]. New York: McGraw-Hill.
- Tan, A., & Nicholson, T. (1997). Flashcards revisited: Training poor readers to read words faster improves their comprehension of text. *Journal of Educational Psychology*, 89, 276–288.
- Taylor, B. M., Pearson, P. D., Peterson, D. S., & Rodríguez, M. C. (2005). The CIERA school change framework: An evidence-based approach to professional development and school reading improvement. *Reading Research Quarterly*, 40, 40–69.
- Torgesen, J. K., Rashotte, C. A., & Alexander, A. W. (2001). Principles of fluency instruction in reading: Relationships with established empirical outcomes. In M. Wolf (Ed.), *Dyslexia, fluency, and the brain* (pp. 333–356). Timonium, MD: York Press.
- Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of Word Reading Efficiency*. Austin, TX: PRO-ED.
- Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Rose, E., Lindamood, P., Conway, T., & Garvan, C. (1999). Preventing reading failure in young children with phonological processing disabilities: Group and individual responses to instruction. *Journal of Educational Psychology*, 91, 579–593.
- Vadasy, P. F., Jenkins, J. R., & Pool, K. (2000). Effects of tutoring in phonological and early reading skills on students at risk for reading disabilities. *Journal of Learning Disabilities*, 33, 579–590.
- Vadasy, P. F., & Sanders, E. A. (in press). Benefits of repeated reading intervention for low-achieving fourth- and fifth-grade students. *Remedial and Special Education*.
- Vadasy, P. F., Sanders, E. A., & Abbott, R. D. (in press). Effects of supplemental early reading intervention at 2-year follow up: Reading skill growth patterns and predictors. *Scientific Studies of Reading*.
- Vadasy, P. F., Sanders, E. A., Jenkins, J. R., & Peyton, J. A. (2002). Timing and intensity of tutoring: A closer look at the conditions for effective early literacy tutoring. *Learning Disabilities Research and Practice*, 17, 227–241.
- Vadasy, P. F., Sanders, E. A., & Peyton, J. A. (2006). Code-oriented instruction for kindergarten students at risk for reading difficulties: A randomized field trial with paraeducator implementers. *Journal of Educational Psychology*, 98, 508–528.
- Venezky, R. L., & Jain, R. (1996). *Tutoring for reading improvement: A background paper*. Unpublished manuscript.
- Wasik, B. A. (1998a). Using volunteers as reading tutors: Guidelines for successful practices. *Reading Teacher*, 51, 562–570.
- Wasik, B. A. (1998b). Volunteer tutoring programs in reading: A review. *Reading Research Quarterly*, 33, 266–292.
- Wiederholt, J. L., & Bryant, B. R. (2001). *Gray Oral Reading Tests-4*. Austin, TX: PRO-ED.
- Wolf, M., & Denckla, M. B. (2005). *RAN/RAS: Rapid Automatized Naming and Rapid Alternating Stimulus tests*. Austin, TX: PRO-ED.
- Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading*, 5, 211–239.
- Woodcock, R. (1998). *Woodcock Reading Mastery Test-Revised/ Normative Update*. Circle Pines, MN: American Guidance Service. (Original work published 1987)
- Yuill, N., & Oakhill, J. (1988). Effects of inference awareness training on poor reading comprehension. *Applied Cognitive Psychology*, 2, 33–45.

(Appendixes follow)



Appendix A

Intercorrelations for Treatment Group

Measure	<i>M</i>	<i>SD</i>	<i>N</i>	1	2	3	4	5	6	7	8	9	10	11
Student assessments														
1. RAN pretest	87.7	10.12	82	—										
2. WR accuracy gain	3.5	5.66	82	.09	—									
3. WR efficiency gain	6.8	7.72	82	.10	.25	—								
4. PRF-U gain	2.9	0.74	82	.32*	.15	.51*	—							
5. PRF-A gain	2.1	0.77	82	.43*	-.07	.25	.46*	—						
6. Fluency rate gain	11.6	9.62	82	.31*	.16	.59*	.68*	-.02	—					
7. Comprehension gain	8.0	13.74	82	-.22	.08	.05	.56*	-.11	-.13	—				
Classroom reading instruction														
8. OTR	15.5	10.34	71	.10	.33*	.41*	.33*	.27*	.41*	-.08	—			
Treatment-related variables														
9. Passage coverage total	95.0	18.22	82	.03	-.05	-.02	.01	-.04	.01	-.17	.07	—		
10. Passage coverage rate	1.9	0.23	82	-.04	.00	-.06	-.11	.02	-.01	-.17	.07	.83*	—	
11. Reading instruction missed <sup>a</sup>	0.8	0.43	73	.11	.04	.12	.17	-.06	.05	-.15	-.05	.25	.16	—

*Note.* RAN = pretest score on the Letter Naming subtest from the Rapid Automatized Naming and Rapid Alternating Stimulus tests; WR accuracy gain = pretest-posttest standard score gain on the Word Identification subtest from the Woodcock Reading Mastery Test—Revised/Normative Update; WR efficiency gain = pretest-posttest standard score gain on the Sight Word subtest from the Test of Word Reading Efficiency; PRF-U and PRF-A = pretest-posttest words-correct-per-minute gains on uniform and alternate passages, respectively, from the Oral Reading Fluency subtest from the Dynamic Indicators of Basic Early Literacy Skills; Fluency rate = pretest-posttest standard score gain on the Rate subtest from Gray Oral Reading Tests-4 (GORT); Comprehension = pretest-posttest standard score gain on GORT Comprehension; OTR = average midyear minutes of classroom literacy spent on oral text reading; Passage coverage total = number of unique passages read; Passage coverage rate = number of unique passages read divided by number of tutoring sessions.

<sup>a</sup> 1 = tutoring occurred during classroom literacy instruction, 0 = otherwise.

\* *p* < .05.

Appendix B

Intercorrelations for Control Group

Measure	<i>M</i>	<i>SD</i>	<i>N</i>	1	2	3	4	5	6	7	8
Student assessments											
1. RAN pretest	90.5	9.99	80	—							
2. WR accuracy gain	1.8	5.29	80	.18	—						
3. WR efficiency gain	8.2	9.25	80	.18	.34*	—					
4. PRF-U gain	2.4	0.87	80	.47*	.31*	.30	—				
5. PRF-A gain	1.5	0.80	80	.26*	.30*	.47*	.31*	—			
6. Fluency rate gain	5.5	8.09	80	.09	.37*	.63*	.57*	.21	—		
7. Comprehension gain	5.4	16.12	80	.13	.22*	.23*	.50*	.31*	.31*	—	
Classroom reading instruction											
8. OTR	13.3	9.32	70	.11	.28*	.37*	.38*	.32*	.35*	.16	—

*Note.* RAN = pretest score on the Letter Naming subtest from the Rapid Automatized Naming and Rapid Alternating Stimulus tests; WR accuracy gain = pretest-posttest standard score gain on the Word Identification subtest from the Woodcock Reading Mastery Test-Revised/Normative Update; WR efficiency gain = pretest-posttest standard score gain on the Sight Word subtest from the Test of Word Reading Efficiency; PRF-U and PRF-A = pretest-posttest words-correct-per-minute gains on uniform and alternate passages from the Oral Reading Fluency subtest from the Dynamic Indicators of Basic Early Literacy Skills; Fluency rate = pretest-posttest standard score gain on the Rate subtest from the Gray Oral Reading Test-4 (GORT); Comprehension = pretest-posttest standard score gain on GORT Comprehension; OTR = average midyear minutes of classroom literacy spent on oral text reading.

\* *p* < .05.

Received November 21, 2006  
Revision received July 31, 2007  
Accepted August 28, 2007 ■

# On Warm Conceptual Change: The Interplay of Text, Epistemological Beliefs, and Topic Interest

Lucia Mason, Monica Gava, and Angela Boldrin  
University of Padua

The aim of this study was to go further than considering only cognitive factors to extend the understanding of the complex, dynamic underlying knowledge revision processes. Fifth graders were assigned to 2 reading conditions. Participants in 1 condition read a refutational text about light, whereas participants in the other read a traditional text. Within each reading condition, students had more or less advanced beliefs about scientific knowledge (complex and evolving vs. simple and certain), as well as high or low topic interest. Overall findings from pretest to immediate and delayed posttests showed that knowledge revision was affected by several interactions among the variables examined. Students who attained the highest scores at both the immediate and delayed posttests were those who had read the refutational text and had high topic interest, as well as more advanced beliefs about scientific knowledge. In particular, the refutational text was more powerful in prompting a restructuring of alternative conceptions about 2 of the 3 light phenomena examined. In addition, students preferred the innovative text to the traditional textbook text.

**Keywords:** conceptual change, refutational text, epistemological beliefs, interest

In their classic article, Pintrich, Marx, and Boyle (1993) called for *hot conceptual change*, that is, to go further than considering only cognitive factors and to also take into account the affective, motivational, and social/contextual aspects that may affect knowledge revision. As described by Sinatra (2005), this article inspired a *warming trend* in theory and research on conceptual change. New models and empirical studies began emphasizing motivation as crucial to knowledge restructuring. Our study examined two learner factors considered by Pintrich and associates (Pintrich, 1999; Pintrich, Marx, & Boyle, 1993) as possible motivational resources for conceptual change: epistemological beliefs and interest. The study also examined an instructional factor, the type of text to be learned to integrate new knowledge into existing conceptual structures. Our study therefore focused on the interplay between instructional and learner variables to extend the understanding of the complex dynamics underlying the process of conceptual change. The investigation of this interplay is in line with most recent research in this field, which highlights the intricacy of the change process and the delicate interactions of multiple factors that shape that process (Alexander & Sinatra, 2007; Murphy & Mason, 2006; Sinatra & Mason, in press). In previous research, these three variables have been examined separately (e.g., Hynd,

2003; Qian & Alvermann, 2000; Qian & Pan, 2002; Venville & Treagust, 1998) or in pairs, for example, refutational text and epistemological beliefs (Mason & Gava, 2007) or epistemological beliefs and interest (Mason & Boscolo, 2004).

## Refutational Text

Most learning required in the school context occurs through successful text reading. Text is still the main medium for acquiring disciplinary knowledge. It is widely accepted that the text comprehension process implies constructing a mental representation of the text or situation model based on integrating information found in the text with the reader's prior knowledge (Chan, Burtis, Scardamalia, & Bereiter, 1992; Goldman & Bisanz, 2002; Kintsch, 1986). This integration often requires readers to restructure their preconceptions if they are incompatible with the new knowledge to be learned. The role of text in promoting conceptual change may therefore be crucial. Science texts are expository texts that provide information about phenomena usually by presenting a series of facts. There is evidence that the structure of standard science texts may induce only superficial processing and may fail to sustain meaningful learning, especially when the reader has alternative conceptions about the topic (e.g., Chambliss, 2002). In the study reported here, we examined the role of the refutational text, which has been found to be effective in science learning because of its structure. A refutational text is a text that acknowledges students' alternative conceptions about a topic, directly refutes them, and introduces scientific conceptions as viable alternatives (Alvermann & Hague, 1989; Hynd, McWhorter, Phares, & Suttles, 1994; Hynd, Qian, Ridgeway, & Pickle, 1991).

The first studies showing the power of a refutational text structure were carried out in the 1980s. For example, Maria and MacGinitie (1987) revealed that fifth and sixth graders more easily abandoned their misconceptions about common topics when they

---

Lucia Mason, Monica Gava, and Angela Boldrin, Department of Developmental and Socialization Psychology, University of Padua, Padua, Italy.

We are grateful to the school principals and teachers for their cooperation in this study. We also thank all of the students involved: Without their work, this study would not have been possible.

Correspondence concerning this article should be addressed to Lucia Mason, Department of Developmental and Socialization Psychology, University of Padua, Via Venezia 8, 35131, Padua, Italy. E-mail: lucia.mason@unipd.it



read a refutational text than when they read an expository one that simply presented the new information. Since then, the effectiveness of refutational texts in knowledge-restructuring processes has been revealed in several studies (e.g., Guzzetti, Snyder, Glass, & Gamas, 1993; Guzzetti, Williams, Skeels, & Wu, 1997; Qian & Pan, 2002). More specifically, the superiority of a refutational text, compared with a traditional one, has been documented in the learning of physics (Hynd, 1998; Wang & Andre, 1991) and biology concepts (Mikkilä-Erdmann, 2002) in students of elementary school (Diakidoy, Kendeou, & Ioannides, 2003), middle school (Mason & Gava, 2007), high school, and college (Alvermann & Hynd, 1989; Chambers & Andre, 1997), as well as in the change of preservice teachers' beliefs about mathematics teaching and learning (Gregoire Gill, Ashton, & Algina, 2004). Through interviews, it also emerged that students prefer refutational texts over other types of science texts about the same topic (e.g., Guzzetti, Hynd, Skeels, & Williams, 1995; Hynd, 2003).

Why are refutational texts powerful? According to Hynd (2003), three explanations can account for the benefits of refutational texts. The first refers to the four conditions for conceptual change, highlighted in the well-known article by Posner, Strike, Hewson, and Gertzog (1982). In other words, refutational texts elicit dissatisfaction with the reader's current conceptions, explain the scientific concept clearly and in depth, make it plausible through believable examples, and finally show the usefulness of the new concept. The second explanation for the power of refutational texts refers to the necessary characteristics of any anomalous data that increase the likelihood that they be accepted as valid. That is, to be accepted, data must be credible, unambiguous, and in multiple forms (Chinn & Brewer, 1993, 1998). The third explanation, grounded in social psychology literature, refers to readers' central, and not peripheral, text information processing, activated by refutational texts. A fourth explanation can be added, regarding the possibility that refutational texts enhance situational interest. As discussed later, this type of interest is triggered by specific text characteristics, for example, concreteness and ease of comprehension (e.g., Schraw, Bruning, & Svoboda, 1995), and may enhance learning.

### Epistemological Beliefs

When confronted with new knowledge to be learned, students do not just activate their preexisting knowledge about the topic but also activate their beliefs about knowledge itself. Beliefs about knowledge and knowing, namely epistemological beliefs, are individuals' representations about the nature, organization, and source of knowledge, its truth value, and the justification criteria of assertions (Hofer & Pintrich, 1997, 2002). In the work of scholars interested in developmental aspects, epistemological thinking has been considered a cognitive structure comprising coherent and integrated representations, which characterizes a level or stage of cognitive development. A shared developmental sequence can be identified in their models (e.g., King & Kitchener, 1994; Kuhn, 2000; Kuhn, Cheney, & Weinstock, 2000). Substantially, the shift indicated is from belief in knowledge as absolute, simple, certain, and transmitted by authorities, to belief in knowledge as relative, idiosyncratic, and uncertain, to belief in knowledge as complex, evolving, rationally evaluated, and derived from reason.

In the work of scholars interested in the relationship between epistemological beliefs and learning processes, the construct of epistemological beliefs has been conceived as multidimensional. Agreement about four epistemological dimensions can be identified in the literature (e.g., Hofer, 2000, 2004; Schommer, 1990, 1994; Schommer-Aikins, 2002; Wood & Kardash, 2002). Two dimensions regard the nature of knowledge. The first concerns beliefs about the simplicity versus the complexity of knowledge: the degree to which knowledge is conceived as compartmentalized or interrelated, ranging from knowledge as made up of discrete and simple facts to knowledge as complex and interrelated concepts. The second dimension concerns beliefs about the certainty versus the complexity of knowledge: the degree to which knowledge is conceived as stable or changing, ranging from absolute to tentative and evolving knowledge. Two belief dimensions regard the nature of knowing. The first concerns the source of knowledge: the relationship between knower and known, ranging from the belief that knowledge resides outside the self and is transmitted to the belief that knowledge is constructed by the self. The second dimension concerns the justification of knowledge: what makes a sufficient knowledge claim, ranging from the belief in observation or authority as a source to the belief in the use of rules of inquiry and evaluation of expertise. According to some scholars (e.g., Schommer, 1990), these four epistemological dimensions may be independent, and an individual may have more sophisticated beliefs in one of them and less sophisticated beliefs in another. According to some other scholars (e.g., Hofer, 2000), beliefs about knowledge and knowing are not a set of independent ideas, but rather are a coherent integration of compatible perspectives, that is, theories, whose core is comprised of the four dimensions. Although divergences exist in their conceptualization of epistemological beliefs (see also Hammer & Elby, 2002), overall it can be said that scholars substantially agree on the essential steps in the development of representations about knowledge and knowing.

In agreement with Bendixen (2002), who examined the perception, description, and solution of epistemic doubt in undergraduate students, it should also be pointed out that epistemological beliefs are not *cold cognition*. The affective nature of deeply held beliefs is an important aspect to be taken into consideration. The recognition of uncertainty and complexity of knowledge can be worrisome for students. Affective consequences (i.e., feelings of fear, confusion, or anxiety) may also derive from abandoning the security of absolutistic stances (Hofer, 2005).

A number of empirical studies have been carried out concerning the relationship between epistemological beliefs and conceptual change learning. To our knowledge, almost all studies are based on the conceptualization of personal epistemology as a multidimensional construct, which has been measured through questionnaires made up of items to be rated on a Likert-type scale. More specifically, original or reduced versions of Schommer's (1990) Epistemological Questionnaire have been used. It should be pointed out, however, that in many cases, the focus has been on two of the four dimensions underlying this instrument, that is, beliefs in the simplicity versus the complexity of knowledge and beliefs in the certainty versus the uncertainty of knowledge. Thus, the label *epistemological beliefs* may refer only to beliefs about the structure (simplicity vs. complexity) and stability (certain vs. evolving) of knowledge. In reviewing previous studies, we specify which dimensions of beliefs are examined.



Qian and Alvermann (1995) found that after reading a text on Newtonian theory, high school students who believed more in the simplicity and certainty of knowledge were less likely to abandon naïve conceptions of motion. In contrast, students who believed more in the complexity and instability of knowledge produced greater change in their conceptual structures. Windschitl and Andre (1998) documented that college students with beliefs in knowledge as complex, evolving, and gradually learned changed their initial conceptions about the cardiovascular system more in the constructivist than in the traditional learning environment. Sinatra and colleagues (Sinatra, Southerland, McConaughy, & Demastes, 2003; Southerland & Sinatra, 2003) investigated not only epistemological beliefs but also cognitive dispositions—such as the disposition to engage in effortful, open-minded thinking and not to identify too strongly with one's beliefs—in relation to the acceptance of human evolution. College students who believed more in complex and uncertain knowledge and had a greater disposition toward critical and open-minded thinking were more likely to accept human evolution than students with less advanced epistemological beliefs who were less disposed toward critical and flexible thinking. Kardash and Scholes (1996) also found that the more college students were disposed to cognitive involvement, the more they wrote paragraphs taking into account the inconclusive nature of controversial evidence presented in the dual-positional text that they read. In addition, the more they believed in uncertain knowledge, the less radical were their initial beliefs about the controversy. In line with Chinn and associates' (Chinn & Brewer, 1993; Chinn & Malhotra, 2002) analysis of the factors that facilitate or constrain theory change, Mason (2000) documented that beliefs about the certainty of knowledge were related to the acceptance of anomalous data, as well as to theory change about a controversial scientific topic. Middle school students who believed in the changing nature of knowledge were more likely to recognize the validity of evidence conflicting with their prior conceptions and, consequently, to change them.

More recently, Nussbaum, Sinatra, and Poliquin (2005) found that college students with evaluativist positions about the physical world—those who believed that different knowledge claims are legitimate but can be rationally compared and evaluated to judge which is more sustainable—showed a greater decline in misconceptions about falling objects than students with absolutist positions. Stathopoulou and Vosniadou (in press) also indicated that 10th graders' physics-related beliefs about the structure, stability, source, and attainability of knowledge were necessary, although insufficient, for a conceptual understanding of physics.

The relationship between epistemological beliefs and text (refutational vs. traditional) reading in conceptual change processes has also been studied (Mason & Gava, 2007). Eighth graders who read a refutational text and believed more in complex and uncertain knowledge changed their misconceptions about biological evolution more than those who read a traditional text.

Why do more sophisticated epistemological beliefs act as a resource whereas less sophisticated beliefs act as a constraint in conceptual change? An account based on the notion of intentionality can be proposed (Sinatra & Pintrich, 2003a). As posited by Sinatra and Pintrich (2003b), a key to intentionality in knowledge revision processes is that "students must have a deliberate *goal orientation* . . . to learn and understand the material" (p. 5). By appealing to this notion, Mason (2003) proposed that beliefs about

knowledge may or may not guide students toward the goal of learning through knowledge revision. In this case, intentionality requires that students recognize that what they already know does not match a new conception or that they perceive the potential of new information. Only beliefs in complex, hypothetical, and evolving knowledge are conducive to that recognition. Once they have recognized that, students must invest their effort in solving problems concerning the current state of their own understanding in order to intentionally produce changes in knowledge (Mason, in press). Systematic or deep processing of the content (Kardash & Howell, 2000) seems to mediate the role of epistemological beliefs in intentional conceptual change. Gregoire Gill, Ashton, and Algina (2004) have documented that preservice teachers who held beliefs in knowledge as simple and certain were less likely to engage in deep thinking about the content of a text and, in turn, were less likely to change their beliefs about mathematics teaching and learning. Epistemological beliefs may also have a direct effect because the conviction that knowledge is simple and certain may lead a student to focus only on items of factual knowledge (Stathopoulou & Vosniadou, in press).

### Topic Interest

We examined a third variable, interest, that may play a role in conceptual change learning. Neglected for decades, this motivational variable has been reconsidered since the mid-1980s (Hidi & Baird, 1986). Conceptualized in terms of both a psychological state and an individual predisposition, interest has been investigated for its influence on cognitive and affective functioning (Renninger, 1990, 2000; Renninger, Ewen, & Lasher, 2002; Renninger & Wozniak, 1985; Schiefele, 1991, 1996). Studies on the relationship between interest and learning have mainly been conducted in the light of the useful distinction between individual and situational interest (Alexander, 1997, 1998; Hidi, 2001; Renninger, 1992; Renninger, Hidi, & Krapp, 1992; Tobias, 1994). Individual interest is distinct from other related motivational constructs and is a relatively stable evaluative orientation toward certain classes of objects, ideas, or events, which is reflected in a relatively enduring predisposition to re-engage with them over time. Situational interest is generated by certain conditions and/or environmental stimuli, such as novelty and intensity. Typically, textual features can elicit a form of interest, called situational interest, that refers to short-term involvement with a class of objects (Renninger, 2000). It has been maintained, however, that situational interest may evolve into individual interest (Hidi & Anderson, 1992), as well as that well-developed individual interest needs to be sustained and challenged (Renninger & Hidi, 2002).

Of special note is research on the development and deepening of interest in subject matter (e.g., Lipstein & Renninger, 2006). A four-phase model has recently been proposed, which relates the constructs of individual interest and situational interest (Hidi & Renninger, 2006). This model includes the following: triggered situational interest, maintained situational interest, emerging individual interest, and well-developed individual interest. Each phase includes both affective and cognitive factors, that is, a certain amount of affect, knowledge, and value. It has been documented that continued (individual) interest in a discipline is related to types of achievement goals. College students' mastery goals pre-



dicted their interest in an introductory psychology course (Harackiewicz, Barron, Tauer, Carter, & Elliot, 2000).

In the current study, we examined topic interest, a motivational construct used in both the individual and situational literatures "to refer to the likelihood of attending to particular subject content . . . or positive feelings for content" (Renninger, 2000, p. 376). We examined topic interest, which is distinct from both individual and situational interest, for two main reasons. First, it is particularly relevant to educators (Ainley, Hidi, & Berndorff, 2002), because this type of interest is most likely evoked when students are introduced to topics in school contexts that they are asked to learn about. Second, for a study focused on changing conceptions about specific scientific phenomena, it was more appropriate to take into account students' degree of interest in the specific topic to be learned about, in addition to their prior knowledge of it.

It is worth noting that topic interest can be considered, to some extent, both individual and situational (Renninger, 2000). Indeed, topic interest can be an expression of individual interest in a specific subject matter, for example, science. At the same time, it can be an expression of situational interest in that it is stimulated by a specific content. Therefore, most topics have both individual and situational sources that contribute to the measured topic interest (Ainley, Hidi, & Berndorff, 2002), especially in young learners, where the former may not be well developed. In the current study, we do conceptually acknowledge that individual interest in science, as well as interest elicited by words or sentences presenting the topic of light, would contribute to the measured topic interest. Nevertheless, topic interest can be conceived as a motivational variable distinct from both individual and situational interests. In addition, although it relates to situational aspects, topic interest can be included as a personal variable, as has been done in most previous investigations. In the current study, it has been considered as a learner variable together with selected relevant dimensions of epistemological beliefs.

Two aspects of topic interest have been distinguished (Krapp, 1999; Schiefele, 1996) and have been taken into account in the study reported here: feeling-related and value-related valences. Feeling-related valences refer to the feelings, for instance, stimulation and involvement, associated with a topic. Value-related valences refer to the attribution of personal significance to a topic. With regard to Hidi and Renninger's (2006) model of interest development, it can be said that topic interest may be an expression of each of the four phases, depending on the contribution of its sources and the degree of cumulative and progressive development. If situational components prevail, one of the two earlier phases is involved, especially in terms of affect or liking and focused attention. If the individual components prevail, one of the later phases is involved, implying not only positive feelings but also stored knowledge and value and repeated engagement.

Although not focused on the change in knowledge, several studies have indicated that topic interest positively influences learning from text. Alexander, Kulikowich, and Schulze (1994) found that college students' topic knowledge and interest significantly predicted comprehension of a technical text on physics. Schiefele (1996) showed that senior high school students with high interest in the topics of prehistoric people and television developed a deeper text comprehension than readers with low interest, regardless of level of prior knowledge. Similarly, in a study by Schiefele and Krapp (1996), topic interest in communication sig-

nificantly affected recall of idea units, elaborations, and main ideas regardless of preexisting knowledge. In Alexander and Murphy's (1998) study, strong interest in concepts such as memory, problem solving, and learned helplessness, as well as willingness to pursue understanding, differentiated highly successful from less successful college students. Boscolo and Mason (2003) found that when high school students' topic interest in global warming was high, it could help their understanding of a text at the superficial level, and also at deeper levels, if associated with high topic knowledge. Ainley, Hidi, and Berndorff (2002) examined whether and how eighth and ninth graders' individual and situational interest contributed to their interest in and learning about different topics, such as body image and X rays. They combined both self-reported measures and online, dynamic measures of affective and cognitive reactions to the reading of science and popular culture texts, through the computer program *Between the Lines*. Topic interest was the outcome of both individual and situational interest factors, which operated in combination or separately. In addition, topic interest was related to affective responses. These were related to persistence that, in turn, was related to learning.

Very few studies have examined the relationship between interest, in either form, and conceptual change, and those that have have revealed that interest may not always be beneficial to conceptual change. In a study of high school students' understanding of genetics, Venville and Treagust (1998) reported some contradictory data. Students interested in the topic produced either high or very low levels of conceptual change. However, students' high interest was in human heredity and not in the microscopic aspects of genetics to be learned. Murphy and Alexander (2004) found that college students with high topic interest were less likely to alter their beliefs. The high correlation between interest and prior knowledge, and the fact that high knowledge often means more resistance to change, could be the cause of the lack of change. As pointed out by Dole and Sinatra (1998), interest could make knowledge revision more difficult if associated with high knowledge. In light of differences in the phases of interest development (Hidi & Renninger, 2006), it is more likely that a well-developed interest acts as a barrier to conceptual change when related to highly developed alternative conceptions.

The positive effect of interest in conceptual change processes has been documented by Andre and Windschitl (2003). They found that college students' interest in learning about electricity affected conceptual change indirectly, through experience that correlated with it, and also directly, independently of its influence on experience. The relationship between topic interest and epistemological understanding has also been examined in high school students' interpretation of a controversy (Mason & Boscolo, 2004). No significant interaction emerged between the two examined variables, although both played a role as single factors.

Why does a high degree of topic interest act as a resource, whereas a low degree may act as a constraint in conceptual change learning? The beneficial effect on interest can be explained by reference to an energizing role in cognitive processing, because it stimulates attentional arousal, positive emotional reactions, effort, and willingness to persist in the task, as documented empirically (e.g., Ainley, Hidi, & Berndorff, 2002) and underlined theoretically (e.g., Alexander, 1998; Hidi, 1990; Renninger, 2000). High interest therefore sustains deep cognitive processing, which is



beneficial to knowledge revision and requires engagement in deep thinking about the content to be learned.

### Research Questions and Hypotheses

To extend current research on conceptual change, we focused on the complex dynamics underlying the process. The aim of this study was to investigate how the text type to be learned and students' epistemological beliefs and topic interest affect revision of their conceptions about a topic. All three variables—each of which can contribute to stimulate and sustain active involvement and deep processing—and their possible relationships in knowledge-restructuring processes were the focus of the study. The variable of students' reading comprehension ability was also considered as a covariate in the statistical analyses. Because the students made up regular heterogeneous classes, their reading ability varied. We chose the topic of light, first because it is a topic central to the Italian science curriculum for the fifth grade, and it had not already been dealt with in the classroom. Second, the students had very low prior scientific knowledge of the topic, and this allowed a better examination of how the chosen variables contributed to new learning.

In particular, the purpose of the study was to examine the following questions:

1. Was students' text retention at the immediate and delayed posttests affected by text type, beliefs about the certainty and development of scientific knowledge, topic interest, and/or their interactions?
2. Was students' overall conceptual change, as reflected by changes in their explanations about light phenomena from pretest to immediate and delayed tests, affected by text type, epistemological beliefs, topic interest, and/or their interactions? In addition, did these factors affect, to the same extent, knowledge restructuring of each of three phenomena about light?
3. Did students prefer a refutational text to a traditional textbook text?

For the first research question, we hypothesized that no significant differences would emerge between participants for the retention of factual information in relation to epistemological beliefs, topic interest, and text type. Because the refutational and traditional texts did not differ in terms of information, participants would retain the same content in the two reading conditions. In addition, as emerged in a previous study (Mason & Gava, 2007), epistemological beliefs would not affect a shallow level of text understanding, which does not require the revision of prior knowledge. Finally, topic interest also would not differentiate text retention because it does not imply deep cognitive processing, which is sustained by the motivational variable.

For the second research question, we hypothesized overall that text type, beliefs about the nature and stability of scientific knowledge, and topic interest would make a difference at the level of deeper text understanding immediately after reading the text and at a delayed posttest. This is the situation model level in Kintsch's (1998) terms; that is, the reader's prior knowledge is integrated with information from the text, which reflects students' conceptual

change. Only when the text information becomes part of the reader's knowledge as it integrates with it, can this information be used in new situations. We hypothesized that the text type would affect students' conceptual change overall. By directly stating and challenging students' alternative conceptions, a refutational text, more than a traditional text, would create or refine students' metacognitive awareness about their own representations and scientific ones, an essential condition for conceptual change (Vosniadou, 2003; Wiser & Amin, 2001). This would increase the likelihood that the former are perceived as limited and the latter as more powerful. We also hypothesized that students with more advanced epistemological beliefs (scientific knowledge as uncertain and changing) would change their alternative conceptions more than those with less advanced epistemological beliefs (scientific knowledge as certain and stable) and that these changes would be persistent.<sup>1</sup> Participants with more advanced epistemological beliefs would be helped in perceiving knowledge problems in their conceptual structures and working toward solving them. Finally, we hypothesized that topic interest would affect knowledge revision overall because this motivational variable would stimulate attentional arousal and deep processing.

For the second research question, we also expected that an interaction among the three variables would emerge. Refutational text reading, more advanced beliefs about the nature of scientific knowledge, and higher topic interest would be a powerful combination of the instructional and personal variables. On the contrary, traditional text reading, less advanced epistemological beliefs, and low interest would lead to a lower performance.

On the basis of the literature, for the third research question we hypothesized that students would prefer a refutational text, which states and challenges their conceptions, to a traditional text, because it is perceived as easier and more helpful in understanding scientific concepts (Guzzetti et al., 1995).

### Method

#### Participants

Initially, the participants involved were 120 fifth graders attending five classes in three public elementary schools in a province of north eastern Italy. All were Caucasian native speakers of Italian and shared a homogeneous middle class social background. Eight of these students were found to have reading difficulties (described later) on administration of a standardized reading comprehension text. In agreement with their teachers, they took part in the study, but their production was not considered in statistical analyses. In addition, a further 6 students were not present at either the pretest or the immediate posttest. There were therefore 106 participants

<sup>1</sup> In the epistemological beliefs literature, terms such as *naïve* and *sophisticated* or *less advanced* and *more advanced* are used. We retain them for reasons of clarity and not to attribute any unfortunate connotation to the terms. In accordance with Schommer-Aikins (2002), we think that when an individual holds sophisticated beliefs, it means that he or she believes mostly in complex and tentative knowledge. However, at the same time, the individual may believe that there is knowledge that is stable and/or isolated. What differentiates an advanced epistemological thinker from a less advanced one is that the belief in certain and simple knowledge is predominant in the latter and is the exception in the former.



who provided data for both the pretest and the immediate posttest. Of these 106 students, 12 were absent on the day of the delayed posttest. Therefore, 94 participants completed the pretest, the immediate posttest, and the delayed posttest (48 girls and 46 boys).

Each class was randomly assigned to one of two reading conditions: a refutational text about light ( $n = 51$ ; 24 girls and 27 boys, including pretest and immediate and delayed posttests) and a traditional text about the same topic ( $n = 43$ ; 24 girls and 19 boys, including pretest and immediate and delayed posttests).

### *Prereading Materials and Tasks*

**Epistemological beliefs.** All but one (Stathopoulou & Vosniadou, in press) of the studies reported in this article used a questionnaire intended to measure general beliefs about knowledge. In this study, we used a more appropriate, science-specific questionnaire, which was an abbreviated version of Conley, Pintrich, Vekiri, and Harrison's (2004) instrument for measuring epistemological beliefs about science in elementary school students. In this way, we tried to overcome the measurement difficulties underlying the use of abstract and domain-general statements. Conley et al.'s (2004) instrument comprises four scales that focus on the core dimensions of epistemological beliefs. Our reduced version comprised 12 items in two scales, one measuring certainty (i.e., "Scientific knowledge is always true") and one measuring development of knowledge (i.e., "Some ideas in science today are different from what scientists used to think"). We used these scales only as they were the most pertinent to the focus of the study (see the Appendix). It should be pointed out that these scales substantially overlap the scales measuring simplicity and certainty in the work of Hofer (2000; Hofer & Pintrich, 1997). Items were rated on a 5-point Likert-type scale (1 = *strongly disagree*, 5 = *strongly agree*). The scale measuring certainty was reversed so that for each item of the two scales, higher scores reflected more sophisticated beliefs.

The alpha reliability coefficient of this reduced version of the instrument was .73. The total score in epistemological beliefs was dichotomized on the basis of the median and was used to create two mutually exclusive groups. One comprised students with more advanced beliefs about the nature of scientific knowledge ( $n = 40$ ; 22 in the experimental condition and 18 in the control condition), and the other comprised students with less advanced beliefs ( $n = 54$ ; 29 in the experimental condition and 25 in the control condition).

**Topic interest.** The authors devised a 10-item interest questionnaire, with items to be rated on a 5-point scale (1 = *not at all*, 5 = *much*), to measure participants' level of interest in the topic of light—the topic of the text to be read (see the Appendix). As in previous studies focused on a single topic to be learned (e.g., Boscolo & Mason, 2003; Mason & Boscolo, 2004; Schiefele, 1996; Schiefele & Krapp, 1996), a questionnaire given before text reading was considered to be a more appropriate way to measure students' evaluative orientation toward content than the simple rating of a title on a content area (e.g., Flowerday, Schraw, & Stevens, 2004). The items were formulated by taking into account the two aspects of interest that have been distinguished (Krapp, 1999; Schiefele, 1996): feeling-related and value-related valences. An example of a feeling-related item is "I would be excited about studying light in science classes." An examples of a value-related

item is "Knowing how we can see the different colors of objects is not important to me." The knowledge component, which has been included in the conceptualization of interest offered by Renninger and other scholars (Renninger, 1990, 1992, 2000; Hidi & Renninger, 2006), was not covered in this questionnaire. We did not expect to find a correlation between knowledge and interest because we assumed that young students would have very little knowledge about scientific topics and would have little understanding of the contribution of knowledge to interest. Nevertheless, the learners could have shown some interest that was based on individual and situational sources and was equated mainly with positive feelings (Renninger, 2000). Furthermore, students' low level of knowledge about the topic would allow a better examination of the role of interest.

Three items (2, 6, and 10) of the questionnaire were reversed so that for each item, higher scores reflected greater interest. The alpha reliability coefficient of this questionnaire was .80. In this case too, the total score for topic interest was dichotomized on the basis of the median and used to create two mutually exclusive groups. One was made up of students with higher topic interest ( $n = 50$ ; 29 in the experimental condition and 21 in the control condition), and the other was made up of students with lower topic interest ( $n = 44$ ; 22 in the experimental condition and 22 in the control condition).

**Prior knowledge.** Eight open-ended questions ascertained participants' preexisting conceptions about light and diffusion (Questions 2, 3, and 7), light and vision (Questions 1, 4, and 5), and light and color (Questions 6 and 8). All questions were of a generative nature (Vosniadou, 1994) and asked students to formulate an explanation for a phenomenon (see the Appendix). Some of these questions were devised by taking into account questions asked in previous studies (Guesne, 1985; Jung, 1986; La Rosa & Mayer, 1991; La Rosa, Mayer, Patrizi, & Vicentini, 1984; Ramadas & Driver, 1989; Ramadas & Shayer, 1987; Watts, 1985) and the grade level of our participants, who were younger than those involved in previous research. Each question was accompanied by a schematic representation of the phenomenon, because much of the teaching of optics in school involves ray diagrams. Diagrams of selected aspects of light-vision phenomena are also very frequent in textbooks discussing light. Some questions also asked the students to add elements to complete a schematic representation. In this case, the elements added and the explanation were considered together, one helping with the understanding of the other. The reliability of open-ended questions was .73 at pretest, .84 at the immediate posttest, and .83 at the delayed posttest.

**Reading comprehension.** Participants were individually administered an Italian standardized test for fifth grade to measure their expository text reading comprehension skills (Cornoldi & Colpo, 1995) and to tease out their correctness and rapidity in decoding. The test consists of a reading text and 10 multiple-choice questions. Students are asked to read the text silently and individually with no time constraints, but time is considered over when 9/10 of the class have finished answering the questions. Students are allowed to refer to the text whenever they need to in order to minimize the memory load. Some of the answers require information explicitly stated in the text and ask respondents to identify characters (animals, entities, etc.), actions, and events, whereas other answers require inferences.



### Reading Materials

The effects of the learning text type on conceptual change were examined by means of two versions of the same text about light. In the traditional text reading condition (control group), students were given a text that combined extracts from common science textbooks (Ajello, Girardet, & Grazzini Hoffman, 1994; Magon, 1996; Saccaro & Signorini, 1996). In the refutational text reading condition (experimental group), students were given a refutational text, which was prepared by modifying the structure of the text used in the former condition. Three parts were added but no other modifications made. Each of these three parts was written to activate, by directly stating and challenging, a specific alternative conception held by the participants. Each part added to the refutational text was written in keeping with the conceptual continuity and text flow. Text coherence was modified as little as possible with the insertion of each part. The three parts did not differ in any aspects of their structure. The first part added addressed the question of the nature and diffusion of light. It was aimed at undermining the conception of a corpuscular nature of light and the conception that the diffusion of light depends on its intensity. The second part added underlined the role of light and its relationship with vision to challenge the naïve conceptions that light alone or light and open eyes are enough to see an object. The third part added was aimed at refuting the idea that color is the property of the object, unrelated to light absorption and reflection phenomena (see the Appendix for excerpts of the traditional and refutational texts).

With these three parts included, the refutational text was much longer than the traditional one. To reduce this discrepancy, the traditional text included additional material that extended and elaborated the information provided. No marginal information was added, which would render the text less focused, therefore disadvantaging students in the control condition. The final version was 481 words, whereas the refutational text was 661 words. Given that the former was still shorter, the control group students were given a further prereading task; that is, they were asked to write down whatever came into their minds on hearing the word *light*. In this way, the amount of time dedicated to the topic was more or less the same for the two groups. Nevertheless, it should be acknowledged that this task, which could activate prior knowledge (i.e., more alternative conceptions about light), could also reduce the likelihood of conceptual change (Alvermann, Smith, & Readence, 1985). However, in this case what the participants wrote, elicited by hearing the word *light*, was not about scientific phenomena that could allow them to activate and focus on alternative representations. The students who had not dealt with light as a scientific topic reported experiences of their everyday life that came to mind. For example, a student's curtains remained open in her bedroom, so she woke up early in the morning before the alarm clock, because there was too much light in the room. Another student described a street that had been dark and was now well lit, which made people happy because they felt more comfortable walking or cycling at night.

### Postreading Materials and Tasks

*Liking the text.* In both reading conditions, students were asked to rate how much they liked the text they had read on a

5-point Likert type scale (1 = *not at all*, 5 = *very much*) and how much they liked to read science texts. These questions were aimed at ascertaining whether there would be a greater preference for the refutational text over the traditional text, which was very similar to any other traditional textbook science text.

*Retention questions.* Five open questions were asked to assess participants' ability to retain facts presented in the text by referring to famous scientists who, in the past, held different views about light. They were challenging to some extent because they were about these views, but they did not ask for any explanations and could be correctly and completely answered, with no inferences, by retaining information that was explicitly expressed in the text. In addition, the answers could be given independently of the students' own conceptions about light. Retaining what a scientist said about light did not imply any reference to their views. Thus, the responses did not represent deep thinking on the content, as did those aimed at measuring conceptual change, which required explanations on the basis of the students' own views (see the Appendix).

*Conceptual change questions.* The same generative open-ended questions used to ascertain participants' prior knowledge were asked again at both immediate and delayed posttests to examine any changes in students' conceptions as well as the stability of these changes. Answers to these questions ascertained text understanding at the level of the situation model (Kintsch, 1998), an understanding based on an integration between prior knowledge and knowledge provided by the text (see the Appendix).

### Coding

Answers to the open-ended questions, which were asked to measure text retention, were scored 0 to 2 according to their degree of correctness and completeness. No points were given for incorrect answers. One point was assigned if the answer consisted of correct but incomplete information. Two points were assigned if the question was answered correctly and completely. For instance, students were asked, "Where does light come from according to Kepler?" The answer "Light rays come from the eyes" scored 0. The answer "Light comes from a lamp" scored 1. The answer "Light rays come from all light sources, for example, the sun or a candle" scored 2. All answers were coded by two independent raters. Each rater read and scored all answers independently. Interrater agreement calculated as a percentage of agreement on the total of the answers was 94%. Disagreements were resolved in conference through discussion in the presence of a third rater.

Answers to the generative open-ended questions asked at pretest and at immediate and delayed posttests to measure conceptual change were analyzed both qualitatively and quantitatively. For the qualitative analysis, different explanation categories were identified for each of the three topics about light. For the quantitative analysis, the answer categories were scored 0 to 2 according to their degree of correctness and completeness, as were the retention questions. For example, one of the questions about vision, taken from Ramadas and Driver (1989), was the following:

A child is in a dark room and cannot see anything. When the lights are turned on in the room she sees a book on a table in front of her. How is she now able to see the book? Explain carefully what is happening



between the book and her eyes. You can draw lines on the diagram to help your explanation. (p. 103)

No points were assigned to unclear or irrelevant answers or to answers expressing alternative conceptions like the following: "The light brightens the book on the table"; "The light brightens up the book. At this point the child can see it. The light illuminates the table well but not all the room"; "When the light is turned on, light rays reach the child's eyes and the book on the table, so she can see it." One point was given for correct but incomplete answers; for example, "The light ray hits the object, so that the child's eye can see it." Two points were given to correct and more complete/elaborated answers; for example, "The rays of the lamp bounce off the book and go to the eyes of the child, who is able to see the book."

All answers were coded by the same two independent raters. Interrater agreement calculated as a percentage of agreement on the total of the answers was 93%. Disagreements were resolved in conference through discussion in the presence of a third rater. Questions are reported in the Appendix.

### Procedure

Data gathering took place in three sessions. The topic of light had not been dealt with previously in any of the students' classes. It is a topic studied in fifth grade, so the teachers' cooperation was extremely valuable. In the first session (pretest), the order of tasks was (a) epistemological beliefs questionnaire, (b) pretest generative open-ended questions, and (c) reading comprehension test. The session took about 2 hr. The second session took place about 1 week after the first. In one condition, participants were asked to study the traditional text, whereas participants in the other condition were asked to study the refutational text. After studying the text (immediate posttest), participants were asked to rate their liking of the text they read and the texts they usually read in their schoolbook, and they were asked to answer the text retention questions and the open-ended generative questions. This session also took about 2 hr. The third session (delayed posttest) took place 2 months after the posttest and lasted about 1.25 hr. Participants were again asked the text retention questions and the open-ended generative questions. Between the immediate and the delayed posttests, the topic of light was not discussed at all in the classrooms to avoid interference from other instructional variables.

### Results

Four *t* tests, controlled for error rates, were performed first to ensure the equivalence of the groups in the two reading conditions, refutational and traditional text, for all the examined measures. The results show that before the reading tasks, there were no statistically significant differences between the two groups of students for any of the measures: epistemological beliefs (refutational:  $M = 37.76$ ,  $SD = 5.78$ ; traditional:  $M = 37.93$ ,  $SD = 5.19$ ),  $t(92) = 0.14$ ,  $p = .88$ ; topic interest (refutational:  $M = 35.55$ ,  $SD = 7.79$ ; traditional:  $M = 36.30$ ,  $SD = 5.99$ ),  $t(92) = 0.51$ ,  $p = .60$ ; prior knowledge (refutational:  $M = 0.29$ ,  $SD = 0.61$ ; traditional:  $M = 0.37$ ,  $SD = 0.92$ ),  $t(92) = 0.48$ ,  $p = .62$ ; and reading comprehensions skills (refutational:  $M = 10.13$ ,  $SD = 2.10$ ; traditional:  $M = 10.51$ ,  $SD = 2.13$ ),  $t(92) = 1.42$ ,  $p = .15$ . As

expected, prior topic knowledge and interest were not correlated ( $r = .01$ ,  $p = .91$ ).

In all the subsequent statistical analyses, the variables of epistemological beliefs and topic interest were dichotomized for two reasons. First, given the nonlinear relation between the dependent and independent variables, the dichotomization would increase the power of the statistical design. Second, the use of the full range value of the variables would have implied a repeated multivariate analysis with one independent variable (type of text) and three covariates (epistemological beliefs, topic interest, and reading comprehension skills), with unclear outcomes. Because we were interested in maximizing the possibility of attaining results that were as clear and interpretable as possible, we opted for the dichotomized variables. This allowed an analysis that offered a clear interpretation of the differences between the two groups.

### Text Retention

A repeated measures analysis of covariance (ANCOVA), with text type (traditional and refutational), epistemological beliefs (less advanced and more advanced), and topic interest (lower and higher) as between-subject variables, with time (immediate posttest and delayed posttest) as the within-subject variable, and with reading comprehension as the covariate, was carried out. It revealed only the effect of the covariate,  $F(1, 85) = 6.25$ ,  $p < .05$ ,  $\eta^2 = .07$ . This means that at both immediate and delayed posttests, reading comprehension skills correlated significantly with the retention of facts introduced in the text. None of the other variables examined affected this more shallow level of text understanding at the two testing times.

### Overall Conceptual Change

Overall changes in explanations about light, vision, and color were measured through a repeated measures ANCOVA with text type (traditional and refutational), epistemological beliefs (less advanced and more advanced), and topic interest (low and high) as between-subject variables, with conceptual knowledge at the three different testing times (pretest, immediate posttest, and delayed posttest) as the within-subject variables, and with reading comprehension as the covariate. From this analysis, the Time  $\times$  Text Type interaction,  $F(2, 84) = 8.31$ ,  $p = .001$ ,  $\eta^2 = .16$ , the Time  $\times$  Topic Interest interaction,  $F(2, 84) = 7.18$ ,  $p = .001$ ,  $\eta^2 = .14$ , and the Time  $\times$  Text Type  $\times$  Topic Interest  $\times$  Epistemological Beliefs interaction,  $F(2, 84) = 4.43$ ,  $p < .05$ ,  $\eta^2 = .09$ , emerged. The analysis also revealed that the covariate, reading comprehension ability, correlated significantly with the three scores,  $F(2, 84) = 9.98$ ,  $p < .001$ ,  $\eta^2 = .19$ .

Regarding the Testing Time  $\times$  Text Type interaction, students who received information from the refutational text were facilitated in conceptual change much more than those who read the traditional text. The former scored higher than the latter at delayed posttest, but their scores decreased more from the immediate posttest to the delayed posttest.

Regarding the Testing Time  $\times$  Topic Interest interaction, participants with high interest outscored students with low motivational involvement at both testing times. Moreover, from the immediate to the delayed posttest, the scores of the highly inter-

ested students decreased more than those of the less interested students, whose scores remained the same.

From the interaction among all three variables, it emerged that at both testing times, the highest scores for conceptual change were obtained by learners with high topic interest, those with more advanced epistemological beliefs, and those who read the refutational text, although these scores decreased significantly from the immediate to the delayed posttest. Scores for students in the experimental condition with low topic interest but more advanced epistemological beliefs increased slightly at the delayed posttest. The same result emerged for the control condition students with low topic interest and less advanced beliefs about the nature of scientific knowledge. The interaction among all three variables can be seen in Figure 1.

Because the refutational text addressed conceptions about three partly connected phenomena, that is, the propagation of light, the relationship between light and vision, and the origin of color, a finer-grained analysis was also carried out to examine in depth the effect of the independent variables on changes in conceptions about each phenomenon. The next sections present the results of this analysis.

### *Conceptual Change About the Nature and Propagation of Light*

A repeated measures ANCOVA with the same between-subject variables and the same covariate, but with the scores of conceptual knowledge about the nature and propagation of light as within-subject variables, revealed a significant Time  $\times$  Text Type interaction,  $F(2, 84) = 10.07, p < .001, \eta^2 = .19$ . Students who read the refutational text changed their conceptions about these aspects of light much more than students who read the traditional text. The scores for their explanations were higher at both testing times (Table 1). Learner variables did not affect conceptual change, except for the covariate of reading comprehension skills,  $F(2, 84) = 7.72, p < .01, \eta^2 = .13$ .

A qualitative analysis identified the conception categories held by students prior to and after text reading (see Table 2). The refutational text was mainly successful in prompting a change in the most common misconception, that is, light as rays that propagate around the light source. It also promoted a change in the alternative conception of light as particles that propagate around the light source. However, not all learners were able to construct scientific knowledge. To illustrate conceptual change due to text reading, which is representative of the data analyzed, we report an example explanation given by a participant (P96, refutational text reading, high topic interest, and more advanced epistemological beliefs) to the third question (see the Appendix). It shows revision of personal conceptions from pretest to posttest.

Pretest: "I have drawn the light which is still and illuminates a specific point."

Immediate posttest: "I have drawn the light rays that do not meet any obstacle so they diffuse in all directions."

### *Conceptual Change About Vision*

A similar repeated measures ANCOVA, with the scores for conceptual knowledge about the relationship of light and vision as within-subject variables, revealed a significant Time  $\times$  Text Type interaction,  $F(2, 84) = 7.77, p = .001, \eta^2 = .15$ , a Time  $\times$  Topic Interest interaction,  $F(2, 84) = 4.06, p < .05, \eta^2 = .09$ , and a Time  $\times$  Text Type  $\times$  Topic Interest  $\times$  Epistemological Beliefs interaction,  $F(2, 84) = 3.41, p < .05, \eta^2 = .07$ . Alternative conceptions about the role of light and its relationship with vision changed more for students who read the refutational text and had high topic interest. At both testing times, the interaction among all three variables showed that more conceptual change was produced by students who read the refutational text and had high topic interest. This interaction also showed that these students had more advanced beliefs about the certainty and development of scientific knowledge (see Table 3). In addition, the ANCOVA revealed a

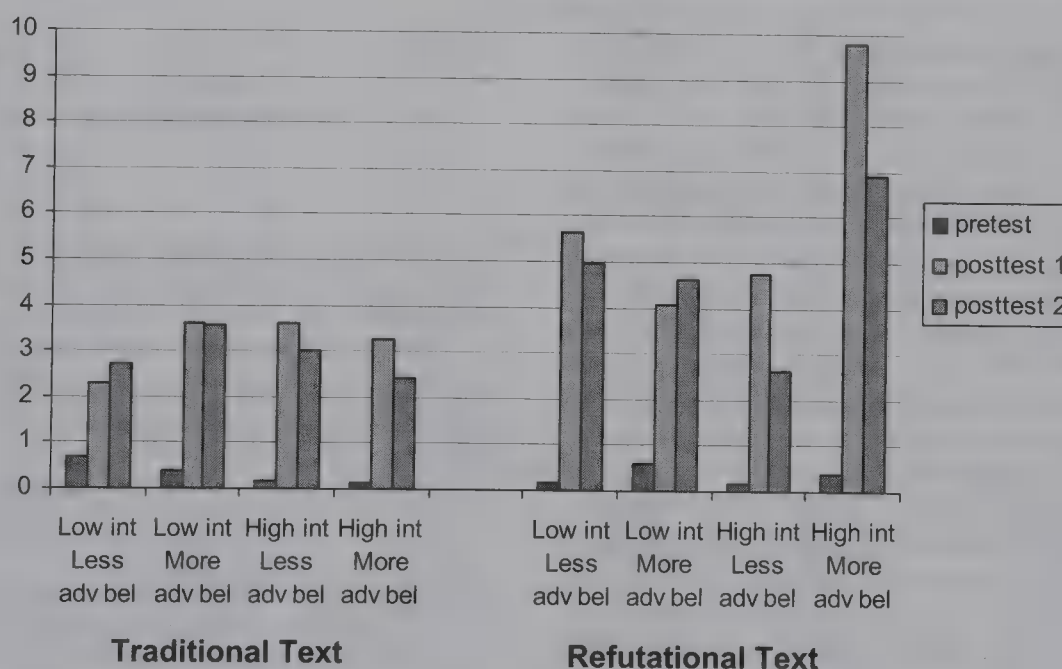


Figure 1. Overall scores for conceptual change showing the interaction between text type, topic interest (int), and epistemological beliefs (bel). adv = advanced.



Table 1  
*Adjusted Marginal Means and Standard Errors of Explanations About the Nature and Diffusion of Light at the Pretest and the Immediate and Delayed Posttests by Text Type*

Text type	n	Pretest		Immediate posttest		Delayed posttest	
		M	SE	M	SE	M	SE
Refutational	51	0.28	0.77	2.73	0.25	2.46	0.27
Traditional	43	0.14	0.83	0.97	0.27	0.99	0.29

Note. Adjustment by covariate of reading comprehension.

significant effect of the covariate, reading comprehension skills,  $F(2, 84) = 9.54, p < .001, \eta^2 = .18$ .

A qualitative analysis identified the conception categories held by students prior to and after text reading (see Table 4). In this case, the refutational text was especially successful in prompting a change in the common, targeted misconception that as well as light, our eyes have an active role in seeing an object. Not all readers, however, were able to learn the new scientific knowledge from the text. The following is a representative example of conceptual change about light and vision as reflected in the explanations given by a participant (P33, refutational text reading, low topic interest, and less advanced epistemological beliefs) to the first question (see the Appendix).

Pretest: "Light illuminates the book on the table and she sees it."

Delayed posttest: "In my opinion the light diffuses in the whole room; a light ray hits the book, it makes the ray bounce off the book, so that it comes to our eyes and we are able to see the colored object."

Conceptual Change About Color

Our last repeated measures ANCOVA, with the scores for conceptual knowledge about color as within-subject variables, revealed only the significant Time  $\times$  Topic Interest interaction,  $F(2, 84) = 3.32, p = .05, \eta^2 = .07$ . Students who reported that they were interested in knowing about light and attributed more importance to the topic were those who changed their naïve conceptions about color more (see Table 5). Only in this analysis did the effect of the covariate, reading comprehension skills, not emerge.

Table 2  
*Frequency and Percentage (in Parentheses) of Responses to Questions About the Nature and Diffusion of Light at the Pretest and the Immediate and Delayed Posttests by Text Type*

Response category and score <sup>a</sup>	Refutational text (n = 51)			Traditional text (n = 43)		
	Pretest	Posttest 1	Posttest 2	Pretest	Posttest 1	Posttest 2
Question 2						
Not pertinent/unclear (0)	6 (11.8)	5 (9.8)	5 (9.8)	9 (20.9)	6 (14.0)	8 (18.6)
Particles around the source (0)	11 (21.6)	3 (5.9)	3 (5.9)	9 (20.9)	4 (9.3)	3 (7.0)
Rays around the source (0)	31 (60.8)	13 (25.5)	17 (33.3)	20 (46.5)	22 (51.2)	23 (53.5)
Particles and infinite rays (0)	2 (3.9)	—	—	2 (4.7)	1 (2.3)	1 (2.3)
Infinite rays (1)	1 (2.0)	6 (11.8)	8 (15.7)	2 (4.7)	5 (11.6)	4 (9.3)
Infinite rays (2)	—	24 (47.1)	18 (35.3)	1 (2.3)	5 (11.6)	4 (9.3)
Question 3						
Not pertinent/unclear (0)	7 (13.7)	5 (9.8)	6 (11.8)	12 (27.9)	9 (20.9)	12 (27.9)
Particles around the source (0)	14 (27.5)	5 (9.8)	10 (19.6)	9 (20.9)	5 (11.6)	4 (9.3)
Rays around the source (0)	17 (33.3)	9 (17.6)	12 (23.5)	16 (37.2)	15 (34.9)	15 (34.9)
Particles and infinite rays (0)	8 (15.7)	—	1 (2.0)	4 (9.3)	2 (4.7)	1 (2.3)
Infinite rays (1)	3 (5.9)	11 (21.6)	5 (9.8)	2 (4.7)	8 (18.6)	4 (9.3)
Infinite rays (2)	2 (3.9)	21 (41.2)	17 (33.3)	—	4 (9.3)	7 (16.3)
Question 7						
Not pertinent/unclear (0)	26 (51.0)	14 (27.4)	14 (27.4)	15 (34.9)	18 (41.9)	18 (41.9)
Particles around the source (0)	3 (5.9)	1 (2.0)	—	1 (2.3)	1 (2.3)	1 (2.3)
Rays around the source (0)	17 (33.3)	13 (25.5)	13 (25.5)	25 (58.1)	17 (39.5)	17 (39.5)
Particles and infinite rays (0)	1 (2.0)	—	—	1 (2.3)	—	—
Infinite rays (1)	3 (5.9)	13 (25.5)	11 (21.6)	1 (2.3)	7 (16.3)	5 (11.6)
Infinite rays (2)	1 (2.0)	10 (19.6)	13 (25.5)	—	—	2 (4.7)

<sup>a</sup> A score of 0 means no points were attributed to the answer; a score of 1 means a correct answer (e.g., "Light expands in all the surrounding area"); a score of 2 means a correct and more elaborated complete answer (e.g., "Light starts from the candle and illuminates until it meets the first obstacle"). See the Appendix for the content of questions.

Table 3  
Adjusted Marginal Means and Standard Errors of Explanations  
About Light and Vision at the Pretest and the Immediate and  
Delayed Posttests by Text Type, Topic Interest, and  
Epistemological Beliefs

Text type, topic interest level, epistemological belief level, and testing time	<i>n</i>	<i>M</i>	<i>SE</i>
Refutational text			
Low topic interest			
Less advanced beliefs			
Pretest	11	0.11	0.10
Posttest 1	11	1.72	0.49
Posttest 2	11	1.28	0.51
More advanced beliefs			
Pretest	11	0.88	0.10
Posttest 1	11	2.37	0.49
Posttest 2	11	1.76	0.50
High topic interest			
Less advanced beliefs			
Pretest	11	0.89	0.10
Posttest 1	11	2.30	0.38
Posttest 2	11	1.20	0.39
More advanced beliefs			
Pretest	18	0.01	0.81
Posttest 1	18	3.94	0.49
Posttest 2	18	2.60	0.50
Traditional text			
Low topic interest			
Less advanced beliefs			
Pretest	10	0.20	0.10
Posttest 1	10	1.55	0.51
Posttest 2	10	1.90	0.53
More advanced beliefs			
Pretest	12	0.33	0.99
Posttest 1	12	1.06	0.47
Posttest 2	12	1.31	0.48
High topic interest			
Less advanced beliefs			
Pretest	8	0.02	0.12
Posttest 1	8	1.19	0.57
Posttest 2	8	0.96	0.59
More advanced beliefs			
Pretest	13	0.13	0.97
Posttest 1	13	1.86	0.45
Posttest 2	13	1.05	0.47

Note. Adjustment by covariate of reading comprehension.

A qualitative analysis identified the conception categories held by students prior to and after text reading (see Table 6). In this case, the refutational text was not powerful in prompting the restructuring of the targeted alternative conception that color is a property of an object. Changes in explanations about color were infrequent and can be illustrated, as an example, by the following participant's answer (P43, refutational text reading, high topic interest, and more advanced epistemological beliefs) to the sixth question (see the Appendix).

Pretest: "I have drawn an arrow to show that the child is looking at the orange. She sees it as orange because its skin is orange."

Delayed posttest: "Light, which is made up of the different colors of the rainbow, absorbs all colors but the orange, so the child sees the orange as orange."

### Liking the Reading Text

For the control group students, a *t* test to compare ratings of how much they liked the text they read revealed no significant difference between the two ratings,  $t(92) = 1.30$ ,  $p = .195$ . Students perceived the text they had just read in the same way as any other science text they were familiar with. On the contrary, for the experimental group, who read the refutational text, a *t* test revealed a significant difference,  $t(92) = 4.16$ ,  $p < .001$ . The unusual text ( $M = 4.24$ ,  $SD = 1.05$ ) just read, which stated and challenged their conceptions, was liked more than the normal textbook science texts ( $M = 3.51$ ,  $SD = 1.17$ ).

### Discussion

This study was aimed at examining the interplay between instructional text and learner variables that may affect conceptual change. Three variables were examined, which had been investigated separately or in pairs in previous research: the structure of the text to be read to gain new knowledge, epistemological beliefs, and topic interest. Two types of instructional texts were used: an ordinary expository science text whose primary function was to give new information and a refutational text that not only gave new, correct information but also explicitly stated and refuted alternative conceptions by presenting the scientific conceptions as viable alternatives. It should be underlined that, as in most previous research on knowledge restructuring, in this study epistemological beliefs reflect more or less advanced representations about only two aspects of scientific knowledge, that is, its nature and its development (simple and certain vs. complex and continuously evolving), and not epistemological beliefs at large. However, these two belief dimensions are the most pertinent to the focus of the study. Topic interest reflects lower or higher affective orientation toward the topic examined.

Our first research question examined whether participants' text retention would be affected by text type, epistemological beliefs, topic interest, and/or their interaction. The findings confirm our hypothesis because none of the examined variables affected the shallow level of text understanding related to retention of factual information. Only the covariate, reading comprehension ability, correlated with the dependent variable.

The second research question examined whether students' overall conceptual change as well as the various aspects of light addressed—which were reflected in their answers to open-ended questions—would be affected by text type, epistemological beliefs, and topic interest. Findings partially confirm our hypotheses. Overall, the positive effect of reading the refutational text confirms previous research (Guzzetti et al., 1993, 1997) and, in particular, confirms an investigation that involved fifth graders learning about energy (Diakidoy et al., 2003), although it did not explicitly address pre- and postreading conceptions about the topic.

Topic interest was also a resource in the knowledge revision process. It must be pointed out that the motivational variable did not correlate with prior knowledge, as emerged from previous studies carried out with much older students (Andre & Windschitl, 2003; Venville & Treagust, 1998). Thus, a higher degree of interest did not mean a higher number of initial alternative conceptions. This could have been because our participants' initial knowledge was generally very limited, so high interest in the topic could be



Table 4  
*Frequency and Percentage (in Parentheses) of Responses to Questions About Light and Visions at the Pretest and the Immediate and Delayed Posttests by Text Type*

Response category and score <sup>a</sup>	Refutational text (n = 51)			Traditional text (n = 43)		
	Pretest	Posttest 1	Posttest 2	Pretest	Posttest 1	Posttest 2
Question 1						
Not pertinent/unclear (0)	1 (2.0)	—	1 (2.0)	1 (2.3)	4 (9.3)	1 (2.3)
Light illuminates the object (0)	20 (39.2)	13 (25.5)	11 (21.6)	21 (48.8)	8 (18.6)	8 (18.6)
Light illuminates the object and the eyes see it (0)	26 (51.0)	7 (13.7)	18 (35.3)	11 (25.6)	11 (25.6)	18 (41.9)
Light illuminates both the object and the eyes (0)	4 (7.8)	1 (2.0)	5 (9.8)	10 (23.3)	6 (14.0)	8 (18.6)
Light goes from the object to the eyes (1)	—	—	—	—	—	—
Light bounces off the object into the eyes (2)	—	30 (58.8)	16 (31.4)	—	14 (32.6)	8 (18.6)
Question 4						
Not pertinent/unclear (0)	16 (31.4)	8 (15.7)	8 (15.7)	13 (30.2)	12 (27.9)	9 (20.9)
Light illuminates the object (0)	4 (7.8)	16 (31.4)	15 (29.4)	4 (9.3)	13 (30.2)	13 (30.2)
Light illuminates the object and the eyes see it (0)	29 (56.9)	10 (19.6)	18 (35.3)	23 (53.5)	15 (34.9)	16 (37.2)
Light illuminates both the object and the eyes (0)	—	—	—	2 (4.7)	—	—
Light goes from the object to the eyes (1)	2 (3.9)	—	—	1 (2.3)	—	—
Light bounces off the object into the eyes (2)	—	17 (33.3)	10 (19.6)	—	3 (7.0)	5 (11.6)
Question 5						
Not pertinent/unclear (0)	8 (15.7)	5 (9.8)	7 (13.7)	9 (20.9)	7 (16.3)	6 (14.0)
Light illuminates the object (0)	38 (74.5)	33 (64.7)	29 (56.9)	30 (69.8)	24 (55.8)	26 (60.5)
Light illuminates the object and the eyes see it (0)	5 (9.8)	3 (5.9)	7 (13.7)	3 (7.0)	6 (14.0)	5 (11.6)
Light illuminates both the object and the eyes (0)	—	1 (2.0)	—	—	—	—
Light goes from the object to the eyes (1)	—	—	—	—	—	—
Light bounces off the object into the eyes (2)	—	9 (17.6)	8 (15.7)	1 (2.3)	6 (14.0)	6 (14.0)

<sup>a</sup> A score of 0 means no points were attributed to the answer; a score of 1 means a correct answer; a score of 2 means a correct and more elaborated, complete answer. See the Appendix for the content of questions.

expressed by students who knew very little about light and not only by those who had high (alternative) prior knowledge. When interest is equated with positive feelings in particular, it may be associated with a low level of content knowledge or little awareness about the contribution of knowledge to interest (e.g., Renninger, 2000).

The main effect of epistemological beliefs did not emerge. However, a significant interaction among convictions about the nature and stability of scientific knowledge, topic interest, and text type, in relation with testing time, has been found generally. As hypothesized, an effective combination of all three factors emerged, although the effect was modest. More conceptual change was produced by students who read the refutational text, those who had more advanced epistemological beliefs, and those with high topic interest. Although when compared with the others these students' performances decreased from the immediate to the delayed posttest, they were still the highest 2 months after reading the

instructional text. This outcome could have been due to the powerful combination of topic interest, which promotes attention arousal, and the other two resources, epistemological beliefs and refutational text. The latter have the potential to help students recognize, more or less directly, the limitations of their own conceptions, to be dissatisfied with them, and to try to make sense of the new knowledge. The advantageous interaction among the three factors may be related to the notion of intentionality in conceptual change (Sinatra & Pintrich, 2003a). These factors could help make learning through conceptual revision a goal to be deliberately pursued (Hynd, 2003; Mason, 2003) to improve current understanding of the phenomena examined.

The interaction among epistemological beliefs, topic interest, and text type sustains deep processing (Alexander et al., 1994). However, processing during text reading (Kardash & Howell, 2000) was not the focus of the current study, and we do not have direct information about whether readers with different epistemo-

Table 5  
*Adjusted Marginal Means and Standard Errors of Explanations About Color at the Pretest and the Immediate and Delayed Posttests by Topic Interest*

Topic interest	n	Pretest		Immediate posttest		Delayed posttest	
		M	SE	M	SE	M	SE
High	50	0.02	0.05	1.02	0.18	0.70	0.18
Low	44	0.15	0.05	0.45	0.18	0.50	0.19

Note. Adjustment by covariate of reading comprehension.

Table 6  
*Frequency and Percentage (in Parentheses) of Responses to Questions About Color at the Pretest and the Immediate and Delayed Posttests by Text Type*

Response category and score <sup>a</sup>	Refutational text (n = 51)			Traditional text (n = 43)		
	Pretest	Posttest 1	Posttest 2	Pretest	Posttest 1	Posttest 2
Question 6						
Not pertinent/unclear (0)	8 (15.7)	16 (31.4)	15 (29.4)	14 (32.6)	13 (30.2)	19 (44.2)
Color is a property (0)	28 (54.9)	12 (23.5)	18 (35.3)	14 (32.6)	11 (25.6)	9 (20.9)
Light gives color (0)	7 (13.7)	10 (19.6)	9 (34.6)	5 (11.6)	6 (14.0)	6 (14.0)
Eyes see colors (0)	5 (9.8)	—	1 (2.0)	4 (9.3)	—	—
Light and eyes allow colors to be seen (0)	3 (5.9)	—	—	4 (9.3)	—	1 (2.3)
Objects reflect color (1)	—	7 (13.7)	3 (5.9)	2 (4.7)	8 (18.6)	3 (7.0)
Light hits the object reflecting color into the eyes (2)	—	6 (11.8)	5 (9.8)	—	5 (11.6)	5 (11.6)
Question 8						
Not pertinent/unclear (0)	7 (13.7)	15 (29.4)	10 (19.6)	10 (23.3)	10 (23.3)	10 (23.3)
Color is property (0)	30 (58.8)	16 (31.4)	18 (35.3)	15 (34.9)	15 (34.9)	15 (34.9)
Light gives color (0)	5 (9.8)	7 (13.7)	9 (17.6)	7 (16.3)	7 (16.3)	8 (18.6)
Eyes see colors (0)	7 (13.7)	1 (2.0)	2 (3.9)	5 (11.6)	2 (4.7)	2 (4.7)
Light and eyes allow colors to be seen (0)	2 (3.9)	1 (2.0)	—	5 (11.6)	—	—
Objects reflect color (1)	—	5 (9.8)	7 (13.7)	1 (2.3)	5 (11.6)	3 (7.0)
Light hits the object reflecting color into the eyes (2)	—	6 (11.8)	5 (9.8)	—	4 (9.3)	5 (11.6)

<sup>a</sup> A score of 0 means no points were attributed to the answer; a score of 1 means a correct answer; a score of 2 means a correct and more elaborated complete answer. See the appendix for the content of questions.

logical beliefs and topic interest process text differently in the two reading conditions. As an alternative explanation, it could be speculated that these students might have—without deep processing—simply noted discrepancies in conceptions. However, this alternative seems less likely in light of the literature on conceptual change, which underlines that it is a demanding process in many respects. In any case, further research is needed to examine the mechanisms through which the beneficial effects of these factors are activated.

In addition, an overall compensation effect of the instructional and motivational variables emerged. Reading an innovative text, which activates and challenges readers' prior knowledge, unlike traditional science texts, compensated for low levels of interest in the topic and less sophisticated convictions about scientific knowledge. Overall, these data confirm a previous study on the relationship between epistemological beliefs and reading a refutational text, from which a powerful combination of the personal and instructional variables also emerged (Mason & Gava, 2007).

The overall outcomes were also found in the finer grained analysis of the change in conceptions about the second phenomenon considered: light in relation to vision. Refutational text and topic interest as separate variables, as well as the interaction between all three variables tested, affected how misconceptions were overcome in favor of the scientific explanation, although not all students abandoned their alternative ideas.

The analysis of conceptual change about the first phenomenon of light addressed, its nature and diffusion, revealed that it was affected only by the type of text read. Specifically, the refutational text succeeded in prompting change in the most commonly held misconceptions, that is, light as rays or particles that propagate only around the source.

A finer grained analysis of conceptual change about the third phenomenon addressed, color, revealed that the refutational text was relatively ineffective in advancing students' scientific knowl-

edge. Only high topic interest helped participants learn more about the scientific conception of the origin of an object's color. The alternative idea of color as a property was the most difficult to change. An explanation for this outcome cannot refer to differences in the refutation provided in that part of the refutational text, because it did not differ in any respect from the previous two parts that were more effective. It can be argued that the scientific conception of the origin of color seems not only as abstract and complex as the others, especially for young students, but also counterintuitive. It is particularly hard to understand that size and shape are properties of an apple but color is not and that its redness must be conceived in relation to the light that the apple reflects. It should be pointed out that in the history of science, the idea that light is not white or yellow but rather contains all colors and the idea that materials absorb some of these colors while reflecting others appeared later than the currently accepted ideas about the nature and propagation of light as well as about vision phenomena.

Furthermore, the correlation between reading comprehension ability and change of conceptions highlights, once again, the importance of this ability in academic learning. Improving students' reading comprehension is, in any case, an essential way to support their conceptual learning from a text.

The overall decreased performance from the immediate to the delayed posttest could be expected, to some extent. The higher decrease in the scores of students with high interest who were in the refutational text condition (see Figure 1) may be explained by a prereading topic interest that was mainly an expression of situational interest elicited by the words and sentences that presented the topic. Topic interest measured in the study could therefore be mainly the expression of a triggered situational interest, that is, a characteristic of the first phase of interest development (Hidi & Renninger, 2006).

The overall increased performance from the immediate to the delayed posttest by students in the experimental condition (those



with more advanced epistemological beliefs and low topic interest) and students in the control condition (those with less advanced epistemological beliefs and low topic interest) cannot be explained by further teaching about light, because the teachers did not deal with this topic during the 2-month period that elapsed between the immediate and delayed posttest. It could be speculated that studying other physics topics may have indirectly helped the students to understand the light phenomena better. However, this hypothesis is rather weak because teachers covered the human cardiovascular and respiratory systems in the interim 2-month period. As an alternative, it could be speculated that students might have had further opportunities to learn about light during the interval by discussing the answers with each other or with their parents and other knowledgeable adults. To some extent, they could have therefore processed the target concepts in the period between the two posttests, perhaps also supported by an increased interest in the topic triggered by the text and questions (Hidi & Renninger, 2006; Renninger et al., 2002).

The third research question asked if students liked the refutational text more than the traditional one. As hypothesized, the refutational text, which helped more in the revision of preexisting conceptions, was perceived as more likeable than the typical textbook text. This outcome is in line with previous research (Guzzetti et al., 1995). Reading a text that made their representations explicit but also highlighted their limitations in favor of scientific ones was a more likeable activity than reading a typical scientific text that transmits knowledge. In this regard, perception of the ease of a learning text could be a key factor in text liking.

Finally, it should also be pointed out that this study examined conceptual change in a complex topic as a consequence of reading one text, and this is a limitation. However, this is what happens in regular classrooms, at least in Italy, from the last years of primary school onward. New knowledge has to be acquired mainly from text and picture reading, especially when abstract concepts are dealt with and the possibilities of experimentation in the classroom are very limited. We agree that conceptual understanding in physics is the product of a gradual and complex process that requires time (Stathopoulou & Vosniadou, 2007). A deeper and more complete understanding of the topic, even at the fifth-grade level, certainly requires more learning activities than the study of one text. Nevertheless, reading a refutational text can be considered a productive starting point for a long and demanding process.

### Conclusion

To conclude, this study documents the role of two learner factors considered in theory by Pintrich and colleagues (Pintrich, 1999; Pintrich et al., 1993) as possible motivational resources for the multifaceted process of conceptual change. These factors, epistemological beliefs and interest, may interact with the instructional material. Higher topic interest was found to be effective in itself and in the interaction with epistemological beliefs and the text type to be learned. A refutational text was also found to facilitate students' understanding of new concepts even when the characteristics of the participants examined were not supportive. This outcome leads us to highlight the potential of refutational texts, especially for the acquisition of concepts that cannot be presented in the classroom through observation and experimentation. The use of refutational texts could also increase students'

situational interest (Hidi, 1990; Krapp, Hidi, & Renninger, 1992; Schraw & Lehman, 2001), which is triggered by specific features of a stimulus or situation, such as text characteristics. In this regard, research has documented that potential situational interest sources are ease of comprehension, text cohesion, concreteness, vividness, engagement, and emotiveness. The most effective sources of situational interest in text learning were found to be ease of comprehension, vividness, and concreteness (Sadoski, 2001; Schraw et al., 1995; Schraw, Flowerday, & Lehman, 2001). It is more likely that a refutational, rather than a traditional, text would be perceived as having these characteristics.

More sophisticated beliefs about the certainty and development of scientific knowledge were found to be effective when these beliefs interacted with all other variables examined. That is, they were best as resources in conjunction with attention arousal, which was stimulated by topic interest, and in relation to reading a text that helped students recognize the limitations of their conceptions and the value of scientific ones. Activities and contexts devised to sustain the development of beliefs about the nature of knowledge are more or less indirect ways to favor the knowledge revision process (Mason, 2002).

In summary, the interplay of factors in text learning that emerged from this study indicate that we have moved beyond cold conceptual change, at least to some extent, and are following the warming trend (Sinatra, 2005). The implications of our findings for conceptual change research highlight that new models of knowledge restructuring must take into consideration the complexity of the process, that is, its multifaceted interactive nature (Sinatra & Mason, *in press*). Models that focus only or mainly on one of the various factors involved in the process cannot account for this complexity or serve as powerful tools for implementing feasible educational interventions aimed at effectively promoting disciplinary knowledge construction and reconstruction in the classroom. It has been recognized that the progress of the research in this multidimensional arena regarding conceptual change may not always be certain and stable (Alexander & Sinatra, 2007). Nevertheless, it is compelling to follow the warming trend (Sinatra, 2005) to increase psychologists' understanding of the complex interplay among key factors underlying conceptual change in the classroom.

From the standpoint of science learning, our findings highlight that if learners hold alternative conceptions, these must be identified in order to tailor instruction to their need for more or less radical knowledge restructuring. The disciplinary material to be learned, for example, the texts to be studied, must be carefully prepared to trigger situational interest. If sustained, this may evolve into a maintained situational interest and gradually into an individual interest (Hidi & Renninger, 2006). This means increasing the likelihood that the new science content is integrated, and not just juxtaposed, into the students' network of conceptions. However, increasing the level of students' understanding of scientific phenomena does not mean that science teaching processes must focus only on conceptual structures, those of the learners, which very often need to be restructured, and those of the disciplinary domain, which are to be acquired. Other noncognitive factors that play an important role in learning processes (motivation in particular) should be taken into consideration. Learners' epistemological representations of the nature of the scientific knowledge they are dealing with, as well as their interest in the content of the knowledge they are asked to understand, are only two of the other factors that affect conceptual change in the classroom.



## References

- Ainley, M., Hidi, S., & Berndorff, D. (2002). Interest, learning, and the psychological processes that mediate their relationship. *Journal of Educational Psychology, 94*, 545–561.
- Ajello, A. M., Girardet, H., & Grazzini Hoffman, C. (1994). *Sussidiario. Quinta classe* [Textbook for fifth grade]. Florence, Italy: La Nuova Italia.
- Alexander, P. A. (1997). Mapping the multidimensional nature of domain learning: The interplay of cognitive, motivational, and strategic forces. In M. L. Maehr & P. R. Pintrich (Eds.), *Advances in motivation and achievement* (Vol. 10, pp. 213–250). Greenwich, CT: JAI Press.
- Alexander, P. A. (1998). The nature of disciplinary and domain learning: The knowledge, interest, and strategic dimensions of learning from subject-matter text. In C. Hynd (Ed.), *Learning from text across conceptual domains* (pp. 263–287). Mahwah, NJ: Erlbaum.
- Alexander, P. A., Kulikowich, J. M., & Schulze, S. K. (1994). The influence of topic knowledge, domain knowledge, and interest on the comprehension of scientific exposition. *Learning and Individual Differences, 6*, 379–397.
- Alexander, P. A., & Murphy, P. K. (1998). Profiling the differences in students' knowledge, interest, and strategic processing. *Journal of Educational Psychology, 90*, 435–447.
- Alexander, P. A., & Sinatra, G. M. (2007). First steps: Scholars' promising movements into a nascent field of inquiry. In S. Vosniadou, A. Baltas, & X. Vamvakoussi (Eds.), *Reframing the conceptual change approach in learning and instruction* (pp. 221–236). Oxford, England: Elsevier.
- Alvermann, D. E., & Hague, S. A. (1989). Comprehension of counterintuitive science text: Effects of prior knowledge and text structure. *Journal of Educational Psychology, 82*, 197–202.
- Alvermann, D. E., & Hynd, C. (1989). Effects of prior knowledge activation modes and text structure on nonscience majors' comprehension of physics. *Journal of Educational Research, 83*, 97–102.
- Alvermann, D. E., Smith, L. C., & Readence, J. E. (1985). Prior knowledge activation and the comprehension of compatible and incompatible text. *Reading Research Quarterly, 20*, 420–436.
- Andre, T., & Windschitl, M. (2003). Interest, epistemological belief, and intentional conceptual change. In G. M. Sinatra & P. R. Pintrich (Eds.), *Intentional conceptual change* (pp. 173–197). Mahwah, NJ: Erlbaum.
- Bendixen, L. D. (2002). A process model of epistemic belief change. In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 191–208). Mahwah, NJ: Erlbaum.
- Boscolo, P., & Mason, L. (2003). Prior knowledge, text coherence, and interest: How they interact in learning from instructional texts. *Journal of Experimental Education, 71*, 126–148.
- Chambers, S. K., & Andre, T. (1997). Gender, prior knowledge, interest, and experience in electricity and conceptual change text manipulations in learning about direct current. *Journal of Research in Science, 34*, 107–123.
- Chambliss, M. J. (2002). The characteristics of well-designed science textbooks. In J. Otero, J. Leön, & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 51–72). Mahwah, NJ: Erlbaum.
- Chan, K. K. C., Burtis, P. J., Scardamalia, M., & Bereiter, C. (1992). Constructive activity in learning from text. *American Educational Research Journal, 29*, 97–118.
- Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science education. *Review of Educational Research, 63*, 1–49.
- Chinn, C. A., & Brewer, W. F. (1998). An empirical text of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching, 35*, 623–654.
- Chinn, C. A., & Malhotra, B. A. (2002). Children's responses to anomalous scientific data: How is conceptual change impeded? *Journal of Educational Psychology, 94*, 327–343.
- Conley, A. M., Pintrich, P. R., Vekiri, J., & Harrison, D. (2004). Changes in epistemological beliefs in elementary science students. *Contemporary Educational Psychology, 29*, 186–204.
- Cornoldi, C., & Colpo, G. (1995). *Prove di lettura MT per la scuola elementare* [New MT tests of reading comprehension for elementary school]. Florence, Italy: Organizzazioni Speciali.
- Diakidoy, I. N., Kendeou, P., & Ioannides, C. (2003). Reading about energy: The effects of text structure in science learning and conceptual change. *Contemporary Educational Psychology, 28*, 335–356.
- Dole, J. A., & Sinatra, G. M. (1998). Reconceptualizing change in the cognitive construction of knowledge. *Educational Psychologist, 33*, 109–128.
- Flowerday, T., Schraw, G., & Stevens, J. (2004). The role of choice and interest in reader engagement. *Journal of Experimental Education, 72*, 93–114.
- Goldman, S. R., & Bisanz, G. L. (2002). Toward a functional analysis of scientific genres: Implications for understanding and learning processes. In J. Otero, J. Leön, & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 19–50). Mahwah, NJ: Erlbaum.
- Gregoire Gill, M., Ashton, P. T., & Algina, J. (2004). Changing preservice teachers' epistemological beliefs about teaching and learning in mathematics: An intervention study. *Contemporary Educational Psychology, 29*, 164–185.
- Guesne, E. (1985). Light. In R. Driver, E. Guesne, & A. Tiberghien (Eds.), *Children's ideas in science* (pp. 10–32). Milton Keynes, England: Open University Press.
- Guzzetti, B. J., Hynd, C., Skeels, S., & Williams, W. (1995). Improving physics texts: Students speak out. *Journal of Reading, 33*, 656–665.
- Guzzetti, B. J., Snyder, T. E., Glass, G. V., & Gamas, W. S. (1993). Promoting conceptual change in science: A comparative meta-analysis of instructional interventions from reading education and science education. *Reading Research Quarterly, 28*, 117–159.
- Guzzetti, B. J., Williams, W. O., Skeels, S. A., & Wu, S. M. (1997). Influence of text structure on learning counterintuitive physics concepts. *Journal of Research in Science Teaching, 34*, 700–719.
- Hammer, D., & Elby, A. (2002). On the form of a personal epistemology. In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology* (pp. 169–190). Mahwah, NJ: Erlbaum.
- Harackiewicz, J. M., Barron, K. E., Tauer, J. M., Carter, S. M., & Elliot, A. J. (2000). Short-term and long-term consequences of achievement goals: Predicting interest and performance over time. *Journal of Educational Psychology, 92*, 316–330.
- Hidi, S. (1990). Interest and its contributions as a mental resource for learning. *Review of Educational Research, 60*, 549–571.
- Hidi, S. (2001). Interest, reading, and learning: Theoretical and practical considerations. *Educational Psychology Review, 13*, 191–209.
- Hidi, S., & Anderson, V. (1992). Situational interest and its impact on reading and expository writing. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 215–238). Hillsdale, NJ: Erlbaum.
- Hidi, S., & Baird, W. (1986). Interestingness—A neglected variable in discourse processing. *Cognitive Science, 10*, 179–194.
- Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist, 41*, 111–127.
- Hofer, B. K. (2000). Dimensionality and disciplinary differences in personal epistemology. *Contemporary Educational Psychology, 25*, 378–405.
- Hofer, B. K. (2004). Epistemological understanding as a metacognitive process: Thinking aloud during online searching. *Educational Psychologist, 39*, 43–55.
- Hofer, B. K. (2005). The legacy and the challenges: Paul Pintrich's contributions to personal epistemology research. *Educational Psychologist, 40*, 95–105.
- Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological



- theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research*, 67, 88–140.
- Hofer, B. K., & Pintrich, P. R. (Eds.). (2002). *Personal epistemology: The psychology of beliefs about knowledge and knowing*. Mahwah, NJ: Erlbaum.
- Hynd, C. (1998). Conceptual change in a high school physics class. In B. Guzzetti & C. Hynd (Eds.), *Perspectives on conceptual change* (pp. 27–36). Mahwah, NJ: Erlbaum.
- Hynd, C. (2003). Conceptual change in response to persuasive messages. In G. M. Sinatra & P. R. Pintrich (Eds.), *Intentional conceptual change* (pp. 291–315). Mahwah, NJ: Erlbaum.
- Hynd, C., McWhorter, Y., Phares, V., & Suttles, W. (1994). The role of instructional variables in conceptual change in high school physics topics. *Journal of Research in Science Teaching*, 31, 933–946.
- Hynd, C., Qian, G., Ridgeway, V., & Pickle, M. (1991). Promoting conceptual change with science texts and discussion. *Journal of Reading*, 34, 596–601.
- Jung, W. (1986). *Considerazioni sulle rappresentazioni mentali degli studenti in ottica* [Considerations on students' mental representations in optics]. *La Fisica nella Scuola*, 19, 150–159.
- Kardash, C. M., & Howell, K. L. (2000). Effects of epistemological beliefs and topic-specific beliefs on undergraduates' cognitive and strategic processing of dual-positional text. *Journal of Educational Psychology*, 92, 524–535.
- Kardash, C. M., & Scholes, R. J. (1996). Effects of preexisting beliefs, epistemological beliefs, and need for cognition on interpretation of controversial issues. *Journal of Educational Psychology*, 88, 260–271.
- King, P. A., & Kitchener, K. S. (1994). *Developing reflective judgment*. San Francisco, CA: Jossey-Bass.
- Kintsch, W. (1986). Learning from text. *Cognition and Instruction*, 3, 87–108.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.
- Krapp, A. (1999). Interest, motivation, and learning: An educational-psychological perspective. *Learning and Instruction*, 14, 23–40.
- Krapp, A., Hidi, S., & Renninger, K. A. (1992). Interest, learning and development. In K. A. Renninger, S. Hidi, & A. Krapp (Ed.), *The role of interest in learning and development* (pp. 3–25). Hillsdale, NJ: Erlbaum.
- Kuhn, D. (2000). Theory of mind, metacognition, and reasoning: A life-span perspective. In P. Mitchell & K. J. Riggs (Eds.), *Children's reasoning and the mind* (pp. 301–326). Hove, UK: Psychology Press.
- Kuhn, D., Cheney, R., & Weinstock, M. (2000). The development of epistemological understanding. *Cognitive Development*, 15, 309–328.
- La Rosa, C., & Mayer, M. (1991). Luce e colore [Light and color]. In N. Gridellini Tomasini & G. Segrè (Eds.), *Scientific knowledge. Students' mental representations* (pp. 185–229). Firenze, Italy: La Nuova Italia.
- La Rosa, C., Mayer, M., Patrizi, P., & Vicentini, M. (1984). Commonsense knowledge in optics: Preliminary results of an investigation on the properties of light. *European Journal of Science Education*, 6, 387–397.
- Lipstein, R., & Renninger, K. A. (2006). "Putting things into words": The development of 12–15-year-old students' interest for writing. In S. Hidi & P. Boscolo (Eds.), *Writing and motivation* (pp. 113–140). Oxford, England: Elsevier.
- Magon, S. (1996). *Per capire. Sussidiario per la classe quinta* [To understand. Textbook for the fifth grade]. Milan, Italy: CETEM.
- Maria, K., & MacGinitie, W. (1987). Learning from texts that refute the reader's prior knowledge. *Reading Research and Instruction*, 26, 222–238.
- Mason, L. (2000). Role of anomalous data and epistemological beliefs in middle students' theory change on two controversial topics. *European Journal of Psychology of Education*, 15, 329–346.
- Mason, L. (2002). Developing epistemological thinking to foster conceptual changes in different domains. In M. Limón & L. Mason (Eds.), *Conceptual change reconsidered. Issues in theory and practice* (pp. 301–336). Dordrecht, the Netherlands: Kluwer Academic.
- Mason, L. (2003). Personal epistemologies and intentional conceptual change. In G. M. Sinatra & P. R. Pintrich (Eds.), *Intentional conceptual change* (pp. 199–236). Mahwah, NJ: Erlbaum.
- Mason, L. (in press). Beliefs about knowledge and revision of knowledge: On the importance of epistemic beliefs for intentional conceptual change in elementary and middle school students. In L. D. Bendixen & F. C. Haerle (Eds.), *Personal epistemology in the classroom: Theory, research, and implications for practice*. New York: Cambridge University Press.
- Mason, L., & Boscolo, P. (2004). Role of epistemological understanding and interest in interpreting a controversy and in topic-specific belief change. *Contemporary Educational Psychology*, 29, 103–128.
- Mason, L., & Gava, M. (2007). Effects of epistemological beliefs and learning text structure on conceptual change. In S. Vosniadou, A. Baltas, & X. Vamvakoussi (Eds.), *Reframing the conceptual change approach in learning and instruction* (pp. 165–196). Oxford, England: Elsevier.
- Mikkilä-Erdmann, M. (2002). Science learning through text: The effect of text design and text comprehension skills on conceptual change. In M. Limón & L. Mason (Eds.), *Reconsidering conceptual change. Issues in theory and practice* (pp. 337–356). Dordrecht, the Netherlands: Kluwer Academic.
- Murphy, P. K., & Alexander, P. A. (2004). Persuasion as a dynamic, multidimensional process: A view of individual and intraindividual differences. *American Educational Research Journal*, 41, 337–363.
- Murphy, P. K., & Mason, L. (2006). Changing knowledge and beliefs. In P. A. Alexander & P. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 305–324). Mahwah, NJ: Erlbaum.
- Nussbaum, E. M., Sinatra, G. M., & Poliquin, A. (2005, April). *The role of epistemological beliefs and scientific argumentation promoting conceptual change*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.
- Pintrich, P. R. (1999). Motivational beliefs as resources for and constraints on conceptual change. In W. Schnotz, S. Vosniadou, & M. Carretero (Eds.), *New perspectives on conceptual change* (pp. 33–50). Amsterdam: Pergamon/Elsevier.
- Pintrich, P. R., Marx, R. W., & Boyle, R. B. (1993). Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational Research*, 63, 167–199.
- Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, 66, 211–227.
- Qian, G., & Alvermann, D. (1995). Role of epistemological beliefs and learned helplessness in secondary school students learning science concepts from text. *Journal of Educational Psychology*, 87, 282–292.
- Qian, G., & Alvermann, D. (2000). The relationship between epistemological beliefs and conceptual change learning. *Reading & Writing Quarterly*, 16, 59–74.
- Qian, G., & Pan, J. (2002). A comparison of epistemological beliefs and learning from science text between American and Chinese high school students. In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology. The psychology of beliefs about knowledge and knowing* (pp. 365–385). Mahwah, NJ: Erlbaum.
- Ramadas, J., & Driver, R. (1989). *Aspects of secondary students' ideas about light*. Leeds, England: Centre for Studies in Science and Mathematics Education.
- Ramadas, J., & Shayer, M. (1987). Schematic representations in optics. In P. Black & A. Lucas (Eds.), *Children's informal ideas: Towards construction of working theories* (pp. 172–189). London: Croom Helm.
- Renninger, K. A. (1990). Children's play interests, representation, and activity. In R. Fivush & J. Hudson (Eds.), *Knowing and remembering in young children* (pp. 127–165). Cambridge, MA: Cambridge University Press.
- Renninger, K. A. (1992). Individual interest and development: Implications

- for theory and practice. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 361–395). Hillsdale, NJ: Erlbaum.
- Renninger, K. A. (2000). Individual interest and its implications for understanding intrinsic motivation. In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 373–404). San Diego, CA: Academic Press.
- Renninger, K. A., Ewen, L., & Lasher, A. K. (2002). Individual interest as context in expository text and mathematical word problems. *Learning and Instruction, 12*, 467–491.
- Renninger, K. A., & Hidi, S. (2002). Student interest and achievement: Developmental issues raised by a case study. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 173–195). New York: Academic Press.
- Renninger, K. A., Hidi, S., & Krapp, A. (Eds.). (1992). *The role of interest in learning and development*. Hillsdale, NJ: Erlbaum.
- Renninger, K. A., & Wozniak, R. H. (1985). Effect of interest on attention shift, recognition, and recall in young children. *Developmental Psychology, 21*, 624–632.
- Saccaro, G., & Signorini, P. (1996). *Otto e trenta. Sussidiario per la classe quinta* [Eight thirty. Textbook for fifth grade]. Bergamo, Italy: ATLAS.
- Sadoski, M. (2001). Resolving the effects of concreteness on interest, comprehension, and learning important ideas from text. *Educational Psychology Review, 13*, 263–281.
- Schiefele, U. (1991). Interest, learning, and motivation. *Educational Psychologist, 26*, 299–324.
- Schiefele, U. (1996). Topic interest, text representation, and quality of experience. *Contemporary Educational Psychology, 21*, 19–42.
- Schiefele, U., & Krapp, A. (1996). Topic interest and free recall of expository text. *Learning and Individual Differences, 8*, 141–160.
- Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *Journal of Educational Psychology, 82*, 498–504.
- Schommer, M. (1994). An emerging conceptualization of epistemological beliefs and their role in learning. In R. Garner & P. A. Alexander (Eds.), *Beliefs about text and instruction with text* (pp. 25–40). Hillsdale, NJ: Erlbaum.
- Schraw, G., Bruning, R., & Svoboda, C. (1995). Sources of situational interest. *Journal of Reading Behavior, 27*, 1–17.
- Schraw, G., Flowerday, T., & Lehman, S. (2001). Increasing situational interest in the classroom. *Educational Psychology Review, 13*, 211–224.
- Schraw, G., & Lehman, S. (2001). Situational interest: A review of the literature and directions for future research. *Educational Psychology Review, 13*, 23–52.
- Sinatra, G. M. (2005). The “warming trend” in conceptual change research: The legacy of Paul R. Pintrich. *Educational Psychologist, 40*, 107–115.
- Sinatra, G. M., & Mason, L. (in press). Beyond knowledge: Learner characteristics influencing conceptual change. In S. Vosniadou (Ed.), *Handbook on conceptual change*. Mahwah, NJ: Erlbaum.
- Sinatra, G. M., & Pintrich, P. R. (Eds.). (2003a). *Intentional conceptual change*. Mahwah, NJ: Erlbaum.
- Sinatra, G. M., & Pintrich, P. R. (Eds.). (2003b). The role of intentions in conceptual change learning. In G. M. Sinatra & P. R. Pintrich (Eds.), *Intentional conceptual change* (pp. 1–18). Mahwah, NJ: Erlbaum.
- Sinatra, G. M., Southerland, S. A., McConaughy, F., & Demastes, J. (2003). Intentions and beliefs in students’ understanding and acceptance of biological evolution. *Journal of Research in Science Teaching, 40*, 510–528.
- Southerland, S. A., & Sinatra, G. M. (2003). Learning about biological evolution: A special case of intentional conceptual change. In G. M. Sinatra & P. R. Pintrich (Eds.), *Intentional conceptual change* (pp. 317–345). Mahwah, NJ: Erlbaum.
- Stathopoulou, C., & Vosniadou, S. (2007). Exploring the relationship between physics-related epistemological beliefs and physics understanding. *Contemporary Educational Psychology, 32*, 255–281.
- Tobias, S. (1994). Interest, prior knowledge, and learning. *Review of Educational Research, 64*, 37–54.
- Venville, G. J., & Treagust, D. F. (1998). Exploring conceptual change in genetics using a multidimensional interpretive framework. *Journal of Research in Science Teaching, 35*, 1031–1055.
- Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction, 4*, 45–69.
- Vosniadou, S. (2003). Exploring the relationships between conceptual change and intentional learning. In G. M. Sinatra & P. R. Pintrich (Eds.), *Intentional conceptual change* (pp. 377–406). Mahwah, NJ: Erlbaum.
- Wang, T., & Andre, T. (1991). Conceptual change text versus traditional text and application questions versus no questions in learning about electricity. *Contemporary Educational Psychology, 16*, 103–116.
- Watts, D. M. (1985). Student conceptions of light: A case study. *Physics Education, 20*, 183–188.
- Windschitl, M., & Andre, T. (1998). Using computer simulations to enhance conceptual change: The roles of constructivist instruction and student epistemological beliefs. *Journal of Research in Science Teaching, 35*, 145–160.
- Wiser, M., & Amin, T. (2001). “Is heat hot?” Inducing conceptual change by integrating everyday and scientific perspectives on thermal phenomena. *Learning and Instruction, 11*, 331–355.
- Wood, P., & Kardash, C. A. (2002). Critical elements in the design and analysis of studies of epistemology. In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 231–260). Mahwah, NJ: Erlbaum.

(Appendix follows)



# Appendix

## Questionnaires, Open-Ended Questions, and Excerpts From the Texts Used in the Study

### The Two Scales Used to Measure Epistemological Beliefs About the Certainty and Development of Science

These two scales are taken from Conley et al.'s (2004, pp. 203–204) questionnaire.

#### *Certainty*

All questions in science have only one right answer.

The most important part of doing science is arriving at the right answer.

Scientists know pretty well everything about science; there is not much more to know.

Scientific knowledge is always true.

Once scientists have the result of an experiment, that becomes the only answer.

Scientists always agree about what is true in science.

#### *Development*

Some ideas in science today are different than what scientists used to think.

The ideas in science books sometimes change.

There are some questions that even scientists cannot answer.

Ideas in science sometimes change.

New discoveries can change what scientists think is true.

Sometimes scientists change their minds about what is true in science.

### The Questionnaire Used to Measure Topic Interest

1. I would be excited about studying light in science classes.
2. I think that there are many more relevant topics than light to learn about in science classes.
3. I think it is important to know what the light we see is made of.
4. If I came across a TV program that talked about light and colors, I would be keen to understand it.
5. I think that during science classes some time could be devoted to talking about light and colors.

6. I am not interested in knowing more scientific aspects of light.
7. I think that light is a difficult but worthwhile topic of science.
8. I would get involved in knowing how we can see the objects around us.
9. I am keen to know what light really is and its characteristics.
10. Knowing how we can see the different colors of objects is not important to me.

### Open-Ended Questions to Measure Text Retention

1. Where does light come from according to Euclid?
2. How does light propagate according to Euclid?
3. Where does light come from according to Kepler?
4. What happens when a light ray hits an object according to Kepler?
5. What did Newton discover about light?

### Open-Ended Questions to Measure Conceptual Change (Deep Text Understanding)

Each question included a diagram, omitted here because of space restrictions.

1. A child is in a dark room and cannot see anything. When the lights are turned on in the room she sees a book on a table in front of her. How is she now able to see the book? Explain carefully what is happening between the book and her eyes. You can draw lines on the diagram to help your explanation (from Ramadas & Driver, 1989, p. 103).
2. It is a dark moonless night. You see a small lamp shining far away. Show where there is light in this drawing. Then, explain why you think that the light is where you drew it (from Ramadas & Driver, 1989, p. 36).
3. In this diagram there is a candle. Draw the light of the candle to show what happens when the candle burns. Explain what you have drawn.
4. Imagine that a person enters a room where there is the candle burning. The person stays in a corner. In your opinion, can the person see the candle? Draw, and then explain what happens.
5. The child in this diagram uses the slide projector to look

at slides of when he was younger. Draw what happens when the projector is turned on. Then, explain how the child is able to see the slides (from Watts, 1985, p. 185).

6. Mark is looking at an orange on the table. Explain why he sees the orange as an orange color.
7. On a clear night a car is parked on a straight street with the lights turned on. A pedestrian who stops on the street can see the car's lights. In the diagram the street is divided into four sections. In which parts is there light? Explain your answer (from La Rosa & Mayer, 1991, p. 197).
8. Marta is looking at the new yellow t-shirt that she got as a present. Explain why she sees it as a yellow color.

#### Excerpt From the Traditional Expository Text

Thanks to Kepler's studies, we now know that light diffuses in a straight line in all directions. Light moves from its source along infinite straight paths called "rays". It continues to diffuse until it meets an object. All bodies that emit light (e.g., the sun and stars)

are called primary light sources. In contrast, the bodies that receive light are called secondary light sources. When the light rays hit an object, they bounce off it and reach our eyes.

#### Excerpt From the Refutational Text

Thanks to Kepler's studies we now know that light diffuses in a straight line in all directions. Some children, however, believe that light diffuses into the environment around the light source only up to a certain point. They believe this because they think that light gets used up as it moves further from the source. If you also think in this way, your conception is not correct. Light moves away from its source along infinite straight paths that are called "rays". It continues to diffuse until it meets an object. The area around the light source appears brighter to us only because there the light rays are closer together, while they broaden as they diffuse in all the surrounding area. When the light rays hit an object, they bounce off it and reach our eyes.

Received April 11, 2006

Revision received July 19, 2007

Accepted July 20, 2007 ■

## ORDER FORM

Start my 2008 subscription to the *Journal of Educational Psychology* ISSN: 0022-0663

\_\_\_\_\_ \$73.00, APA MEMBER/AFFILIATE \_\_\_\_\_  
 \_\_\_\_\_ \$161.00, INDIVIDUAL NONMEMBER \_\_\_\_\_  
 \_\_\_\_\_ \$450.00, INSTITUTION \_\_\_\_\_  
*In DC add 5.75% / In MD add 6% sales tax*  
**TOTAL AMOUNT ENCLOSED** \$ \_\_\_\_\_

**Subscription orders must be prepaid.** (Subscriptions are on a calendar year basis only.) Allow 4-6 weeks for delivery of the first issue. Call for international subscription rates.



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

#### SEND THIS ORDER FORM TO:

American Psychological Association  
Subscriptions  
750 First Street, NE  
Washington, DC 20002-4242

Or call **800-374-2721**, fax **202-336-5568**.  
TDD/TTY **202-336-6123**.

For subscription information, e-mail:  
**subscriptions@apa.org**

☐ **Check enclosed** (make payable to APA)

**Charge my:** ☐ VISA ☐ MasterCard ☐ American Express

Cardholder Name \_\_\_\_\_

Card No. \_\_\_\_\_ Exp. Date \_\_\_\_\_

\_\_\_\_\_  
Signature (Required for Charge)

#### BILLING ADDRESS:

Street \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

Daytime Phone \_\_\_\_\_

E-mail \_\_\_\_\_

#### MAIL TO:

Name \_\_\_\_\_

Address \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

APA Member # \_\_\_\_\_ EDUA08



# Relationships of Three Components of Reading Fluency to Reading Comprehension

Susan Lutz Klauda and John T. Guthrie  
University of Maryland, College Park

This study examined the relationships of 3 levels of reading fluency—the individual word, the syntactic unit, and the whole passage—to reading comprehension among 278 5th graders heterogeneous in reading ability. Hierarchical regression analyses revealed that reading fluency at each level related uniquely to performance on a standardized reading comprehension test in a model including inferencing skill and background knowledge. The study supports an automaticity effect for word recognition speed and an automaticity-like effect related to syntactic processing skill. In addition, hierarchical regressions using longitudinal data suggest that fluency and reading comprehension have a bidirectional relationship. The discussion emphasizes the theoretical expansion of reading fluency to 3 levels of cognitive processes and the relations of these processes to reading comprehension.

**Keywords:** reading comprehension, reading fluency, prosody, word recognition speed, syntactic processing

Recently, there has been increased attention to reading fluency. The individual constituents of fluency (Schwanenflugel, Hamilton, Kuhn, Wisenbaker, & Stahl, 2004; Wolf & Katzir-Cohen, 2001) and the relationship of fluency to comprehension (Jenkins, Fuchs, van den Broek, Espin, & Deno, 2003b; Kuhn & Stahl, 2003) have been of particular interest.

Although speed and accuracy in reading have long been considered hallmark components of fluency (Fuchs, Fuchs, Hosp, & Jenkins, 2001), appropriate expression is now considered an additional defining component (Kuhn & Stahl, 2003; National Reading Panel, 2000). Appropriate expression refers to using prosodic features of language, such as emphasis, pitch changes, pause placement and duration, and phrasing in accord with syntactic structure so that text is translated aloud with the tonal and rhythmic characteristics of everyday speech (Dowhower, 1991; Kuhn & Stahl, 2003). In the present study, we abide by this definition, while recognizing that the terms *expressiveness* and *prosody* are occasionally used differentially (e.g., Cowie, Douglas-Cowie, & Wichmann, 2002).

It is also useful to consider fluency in terms of the text units, or levels, at which one is fluent. However, as stated by Wolf and Katzir-Cohen (2001), “Few current approaches attempt to define fluency in terms of either its component parts or its various levels of reading subskills—that is, letter, letter pattern, word, sentence, and passage” (p. 218). In the present study, we investigated fluency at the latter three of these levels, with emphasis on speed and accuracy at the word level and aspects of prosody at the sentence and passage levels.

## Relationship Between Fluency and Comprehension

Empirical studies of the relationships among reading skills have often reported moderate to high positive correlations between measures of fluency and comprehension. These correlations appear in research involving students from elementary through high school. Moreover, the studies represent large, diverse samples of students and smaller, focused samples, such as middle-school children with reading disabilities, and have used a variety of measures of fluency and comprehension (e.g., Daane, Campbell, Grigg, Goodman, & Oranje, 2005; Fuchs, Fuchs, & Maxwell, 1988; Jenkins et al., 2003b; Pinnell, Pikulski, Wixson, Campbell, Gough, & Beatty, 1995; Rasinski et al., 2005; Yovanoff, Duesbery, Alonzo, & Tindal, 2005). In addition, interventions that focus on increasing fluency have often been associated with significant gains in both fluency and comprehension, as well as correlations between the gains in each skill, for new readers and for struggling readers through high school (see reviews by Chard, Vaughn, & Tyler, 2002; Kuhn & Stahl, 2003; National Reading Panel, 2000).

Some studies, however, have indicated dissociations between fluency and comprehension when fluency is defined as accuracy or speed in reading individual words or pseudowords. For example, Oakhill, Cain, and Bryant (2003) found that different sets of component skills predicted word recognition and comprehension in average 7- to 9-year-old readers. However, a measure of syntactic processing predicted variance in both word recognition and

---

Susan Lutz Klauda and John T. Guthrie, Department of Human Development, University of Maryland, College Park.

The research described herein was presented in part at the 2007 American Educational Research Association Meeting, Chicago. This work was supported by Interagency Educational Research Initiative Award 0089225 as administered by the National Science Foundation. The findings and opinions expressed here do not necessarily reflect the position or policies of the Interagency Educational Research Initiative, the National Science Foundation, or the University of Maryland.

We thank A. Laurel Hoa, Ellen Kaplan, and Jenna Ogilvie for assisting in the development of the reading fluency rubric and data coding.

Correspondence concerning this article should be addressed to Susan Lutz Klauda, Department of Human Development, University of Maryland, 3304 Benjamin Building, College Park, MD 20742. E-mail: susan3@umd.edu

comprehension. In addition, Jackson and Doellinger (2002) identified six university students from a group of 17 poor recoders who comprehended as well as their peers with average recoding and comprehension skills, thereby suggesting that some poor recoders develop compensatory mechanisms that aid their comprehension. Similarly, Walczyk, Marsiglia, Johns, and Bryan (2004) found that when third graders read aloud under unrestricted circumstances (i.e., without time pressure, after listening to a model who used compensation techniques, and with the knowledge that their comprehension of what they read would be tested), literal comprehension did not correlate with accuracy in reading a list of words unrelated to the comprehension texts.

### Explaining Links Between Fluency and Comprehension

There are two major theoretical views of the processes by which fluency and comprehension may be related, one focusing on fluency at the word level and the other on fluency at the syntactic level, or fluency in reading sentences and phrases. Moreover, consideration of fluency at a third level—that of the passage as a whole—may add to our understanding of the links between fluency and comprehension. Specifically, we investigated whether fluency at the passage level, based on processing of passage-level features, or understanding of the macrostructure of the text (Kintsch & Kintsch, 2005) would add to the prediction of comprehension beyond fluency at the word and syntactic levels.

#### *Word Level*

According to automaticity theory (LaBerge & Samuels, 1974) and verbal efficiency theory (Perfetti, 1985), growth in fluency, specifically in the speed component, facilitates reading comprehension. That is, as word recognition becomes faster, it eventually becomes automatic, allowing the attention that was once required for the task of word decoding to be devoted to comprehension. Correlations between word reading speed, for words presented in list as well as passage form, and comprehension provide support for this view (e.g., Fuchs, Fuchs, & Maxwell; Jenkins et al., 2003b; McCormick & Samuels, 1979; Perfetti & Hogaboam, 1975). Furthermore, Schwanenflugel et al. (2004) found that reading with prosody did not contribute above and beyond efficiency of word recognition to the prediction of comprehension, suggesting that automaticity theory alone might be sufficient to explain the relationship between fluency and comprehension. An alternative to automaticity theory to account for relationships between word reading speed and comprehension is interactive theory. According to Rumelhart (1994) and others, top-down processes of language comprehension (e.g., syntactic knowledge) facilitate word recognition speed. We do not attempt to distinguish between these theories in this study.

#### *Syntactic Level*

Theoretically, comprehension and fluency may also be related because they share a basis in some of the semantic and syntactic processes involved in processing language at the phrase or sentence level (Jenkins et al., 2003b). Although Jenkins et al. (2003b) focused primarily on the semantic processes, we focus more on the

role of syntactic processing of sentences as an individual constituent of fluency and contributor to comprehension.

Support for the view that syntactic processing is an important element of fluency comes from work by Rasinski (1985), revealing that the ability to parse text into meaningful phrases mediated the relationship between word recognition and fluency development in good and average third- through fifth-grade readers. In addition, Berninger, Abbott, Billingsley, and Nagy (2001) identified a subtype of disabled readers characterized by errors in executive coordination at the sentence level, including inattention to syntax (represented by misordering words) and failure to read with prosody. Furthermore, as posited by Schreiber (1980) and substantiated by Dowhower (1987), the intervention method of repeated readings (Samuels, 1979) appears to lead to gains in fluency and comprehension, not because of the practice it affords for word recognition but at least partly because it forces readers to learn to compensate for the lack of prosodic cues in written text by instead using syntactic as well as semantic, morphological, and contextual cues.

Further evidence, according to both Young and Bowers (1995) and Kuhn and Stahl (2003), that syntactic processing contributes to comprehension comes from two categories of studies: (a) those showing that training readers to segment text into meaningful phrases or giving them text that has already been segmented results in better comprehension (e.g., Amble & Kelly, 1970; Cromer, 1970; Mason & Kendall, 1979; O'Shea & Sindelar, 1983) and (b) those showing that teaching children to identify words faster, although associated with gains in decoding speed, is not linked with significant gains in comprehension of text composed of the trained words (e.g., Fleisher, Jenkins, & Pany, 1979; Grant & Standing, 1989; Spring, Blunden, & Gatheral, 1981).

#### *Passage Level*

In addition to sharing a basis in processes at the phrase and sentence level, fluency and comprehension may depend on processes relevant to the passage as a connected whole. That is, to some degree, performance on fluency and comprehension tasks that involve reading passages (either silently or out loud) may both be reliant on processing of features that lend the text an overall organization and coherence or, as termed by Kintsch and Kintsch (2005), the macrostructure of the text.

For narrative passages, these features may include categories of story grammar, such as those formulated by Stein and Glenn (Stein, 1979; Stein & Glenn, 1979) for a simple story: setting, initiating event, internal response, attempt, consequence, and a reaction. Empirical research has provided clear evidence that understanding story grammar aids reading comprehension (see Olson & Gee's review, 1988), and we posit that an understanding of story grammar might also assist reading fluency, namely the prosodic or expressive aspects of it. In other words, when a reader encounters a section of a passage that fits into one of the story grammar categories, the reader will mark that section with appropriate prosodic features if the reader has developed awareness of that category of story grammar.

For expository passages, developing an understanding of the macrostructure involves discerning the overall purpose or basic organizational structure of the text, which, according to Meyer, Young, and Bartlett (1989), may be causation, description, se-



quence, problem solution, or comparison. Each structure is associated with signaling words. For example, *as a result*, *because*, and *in order to* indicate the causation structure. Meyer and colleagues (e.g., Meyer, 1999; Meyer et al., 2002, 1989) have found clear links between identification of expository text structures and comprehension for both children and adults, but as with narrative passages, apparently no studies have investigated whether awareness of the text structure is also associated with fluency.

The highest levels of rubrics that have been used to assess the quality of children's passage oral reading place some emphasis on the expressiveness dimension of fluency, implying that the best readers have a sense of the passage's macrostructure. For example, to be rated at the highest level of the 1 to 4 scoring rubric used in the studies of oral reading fluency conducted by the National Assessment of Educational Progress (Daane et al., 2005; Pinnell et al., 1995), the reading must meet the criteria that "some or most of the story is read with expressive interpretation" (Daane et al., 2005, p. 28). To distinguish the potential differential relations of syntactic understanding at the sentence level and expressiveness at the passage level with comprehension, these variables were rated independently in the current study.

### Mediators of the Relationship Between Fluency and Comprehension

Automaticity theory (LaBerge & Samuels, 1974) posits that automatic word recognition frees resources for higher level processes involved in comprehension. However, few researchers have endeavored to determine exactly which processes are given attention once automaticity in word recognition is achieved. In line with Kintsch (1988, 1998); Graesser, Singer, and Trabasso (1994); and Thurlow and van den Broek (1997), we view inferencing and the integration of one's knowledge base with the text that one is reading as key higher order processes involved in comprehension. Thus, we also investigated whether these were indeed processes that became more active with growth in fluency or, in other words, whether they mediated the relationship between fluency and comprehension.

### Directionality of the Relationship Between Fluency and Comprehension

The direction of causality between fluency and comprehension is currently a matter of some debate (Wolf & Katzir-Cohen, 2001). As pointed out by Stecker, Roser, and Martinez (1998), there is evidence that fluency is both a contributor to and a product of comprehension; they thus advocated viewing comprehension and fluency as having a reciprocal causal relationship, a view currently espoused by practitioners as well as reading researchers (Pikulski & Chard, 2005). Traditionally, however, researchers have theorized that fluency primarily facilitates comprehension, in line with automaticity theory (LaBerge & Samuels, 1974). In contrast, others such as Kuhn and Stahl (2003) and Young and Bowers (1995) have contended that appropriate application of prosodic features, along with speed in word recognition, plays an important role in facilitating comprehension.

Others view the relationship between fluency and comprehension as varying to some degree in relation to one's reading skill. Jenkins et al.'s (2003a) findings indicated, for example, that com-

prehension may especially facilitate fluency for children of higher reading ability, whereas weak word recognition skills may be what limit the fluency and comprehension development of poor readers. Furthermore, as described by Fuchs et al. (2001), understanding text and relating it to prior knowledge in the domain of the text may help readers correctly anticipate words in connected text that they might struggle with if they were presented out of context. As suggested by Young, Bowers, and MacKinnon (1996), this process may play a particular role in poor readers' gains in reading speed, freeing resources for comprehension and permitting proper phrasing and expressiveness to emerge as by-products of the dual improvements in word reading efficiency and comprehension. Weighing in on the debate about the nature of the relationship between fluency and comprehension, Paris, Carpenter, Paris, and Hamilton (2005) cautioned against viewing fluency as necessary or sufficient for comprehension because of the multiple shared processes that may account for relations between fluency and comprehension, such as vocabulary, syntactic knowledge, and background knowledge.

### Research Questions

The present study addressed four questions regarding the relationship between reading fluency and comprehension:

1. To what extent does each type of fluency—word, syntactic, and passage—correlate with reading comprehension when the other types of fluency, inferencing, and prior knowledge are statistically controlled?
2. To what extent is the association of each type of fluency with reading comprehension mediated by inferencing and prior knowledge?
3. To what extent does each type of fluency correlate with reading comprehension when the other types of fluency are statistically controlled?
4. To what extent does fluency predict changes in comprehension over a 12-week period, and to what extent does reading comprehension predict changes in fluency over the same length of time?

### Method

#### Participants

The participants in the present study were 278 fifth-grade students from 13 classrooms in three schools. The schools were located in a small city in a mid-Atlantic state. As shown in Table 1, the sample was representative of the school district in terms of the percentage of male and female students and the percentages of students receiving English as a Second Language (ESL) instruction, special education services, and free and reduced-price meals (FARMS). The students' mean grade equivalency score on the Gates-MacGinitie Reading Test, taken near the beginning of the school year, was 5.87 ( $SD = 3.17$ ). It should be noted that the sample included students reading several years below to several years above grade level.

Table 1  
*Demographic Characteristics of Students in the Sample and in the School District*

Characteristic	Sample (%)	District (%)
Gender		
Male	49.6	51.2
Female	50.4	48.8
Ethnicity		
African American	20.5	10.8
Asian	2.9	3.8
Caucasian	65.2	79.1
Hispanic	8.8	6.0
Other	2.6	0.3
ESL	4.8	5.4
Special education	9.2	11.6
FARMS		18.1

*Note.* The district percentages for gender and ethnicity represent students at the elementary school, middle school, and high school levels combined; the English as a Second Language (ESL), special education, and free and reduced-price meals (FARMS) represent elementary school students only.

### Measures

*Rationale for measurement design.* Three measures of fluency and single measures of reading comprehension, inferencing, and background knowledge were used. As detailed below, the reading comprehension and inferencing assessments, as well as two of three fluency measures, shared a partially common text base. Although this methodological approach may raise questions about measurement independence, it enables specification of the conditions under which different aspects of fluency relate to comprehension. With this approach, we expected the variables under study to show strong relationships, or at least relationships stronger than those that would be obtained if the measures were based on entirely different texts. In this way, then, text characteristics were controlled across the reading tasks, which could otherwise differentially influence the construct being measured.

*Gates-MacGinitie Reading Test (GMRT).* GMRT Forms S and T (comprehension tests) were used. This test was administered to all students in late August and mid-December. At the first test point, students were assigned to Level 4, 5, or 6 of Form S of the test on the basis of teacher judgments of their general reading ability and school records. At the second test point, students were reassigned to the same test levels of Form T unless they had scored below or above a predetermined cutoff point at the first test point. The extended scale scores were used in analyses because the raw scores are subject to skew. Also, the extended scale scores fulfill the assumption of equal interval scaling more adequately than raw scores because they normalize the raw scores, especially at the ends of the continuum. Furthermore, the meaning of extended scale scores is the same, regardless of the test level administered (MacGinitie, MacGinitie, Maria, & Dreyer, 2000).

*Inferencing assessment.* Our inferencing measure was designed to assess the skill of readers in recognizing implicit information in text. That is, an inference is a connection made between elements of a text based on the text itself and the reader's background knowledge; this definition is similar to the definition of bridging inferences used by Hannon and Daneman (2001) and Kintsch (1998). In addition, like Hannon and Daneman, true-false

items based on the stimulus texts were used to assess inferencing skill.

Three levels of the inferencing assessment were created to correspond with the three levels of the GMRT that were used at Time 1. From the first half of each level-form combination of the GMRT comprehension subtest, two passages (one narrative and one expository) were selected. These passages, which are referred to below as the *stimulus texts*, provided the text base for the items on the inferencing assessment and for two of our fluency measures (the Word Recognition Assessment and the Passage Oral Reading Assessment). Students received inferencing and fluency tests that corresponded with the level of the comprehension test they received.

On each level of the inferencing assessment, the two stimulus texts from the GMRT were presented, each followed by six true-false items. One item in each set of six was constructed to measure literal comprehension or very low level inferencing, whereas the other five items were constructed to measure middle- to high-level inferencing. For instance, an item could require processing only one or two sentences, a single paragraph, or the entire passage. Before completing the assessment, students completed two practice items that were based on a sample passage. Cronbach's alpha was .52.

*Background knowledge assessment.* The background knowledge measure assessed knowledge of ecological concepts. This measure was considered a proxy for a test of more general knowledge, based on the assumption that knowledge of one topic area would predict knowledge of other topics. Nineteen multiple-choice items were constructed; each item consisted of a stem and four alternative words or phrases to complete the stem. The Cronbach's alpha value obtained for this measure was .63. In addition, a correlation of .71 was observed for test-retest scores from a quarter of the sample, who took the test again 3 months later.

*Word recognition assessment.* This assessment measured fluency at the word level, defined in this study as how quickly students could correctly identify individual words presented in a list. Two word lists were created that corresponded with each stimulus text by first placing all unique words from each text in order by length, with the exception that proper nouns were placed at the end of each list. The full list was then divided in half, into A and B forms, by alternately placing the ordered words on separate lists. Students received either the A or the B list that corresponded to one of the stimulus passages for their assigned level, using a counterbalancing system detailed below in the *Procedure* section. The lists varied in length from 28 to 44 words.

The directions for the Word Recognition Assessment (WRA), which was individually administered by research team members, were as follows:

I'm going to give you a list of words to read out loud. You'll begin at the top left of the list, and read down each column. You may use your finger to help you keep your place if you would like. Read the words as quickly as you can without making mistakes. If you come to a word that you don't know, skip it and go to the next word. Continue reading until I ask you to stop or until you finish the entire list. I'm going to use a stopwatch to see how long it takes you. Do you have any questions?

After reading these directions aloud, the administrator gave the student a practice list consisting of six words and then the test list.



During the testing phase, the administrator marked any words that the student read incorrectly or omitted and used a stopwatch to record the number of seconds that it took the student to read the full list. The WRA was scored by calculating the number of words read correctly per minute, termed *word reading speed* in the present study. Moreover, word accuracy was determined by calculating the percentage of words each student read correctly. However, consideration of the high mean for percentage correct (91.6), in combination with the fact that the standard deviation for words correct per minute was nearly twice that for percentage correct (26.0 vs. 13.5) led to the decision to use only words correct per minute in the analyses.

Reliability for the WRA was determined by calculating the test-retest correlation for word reading speed for a quarter of the original sample, who were retested 3 months later with word lists based on different passages. This test-retest correlation was .72.

*Woodcock-Johnson III Reading Fluency Test.* Performance on the Reading Fluency Test from the Woodcock-Johnson III (WJ-III) Diagnostic Reading Battery measured fluency at the syntactic level, defined in this study as accuracy and speed in processing phrase and sentence units of text. This test consisted of 98 simple sentences primarily describing common animals and objects. Students were directed to silently read as many of the sentences as they could within 3 min, circling *Y* for "yes" or *N* for "no" after each sentence, depending on whether it was true or false. Scores on the test equal the number of correct responses minus the number of incorrect responses. The raw score associated with a grade equivalency of 5.0 is 43. The internal consistency coefficient for age 10 is .90, and the 1-year test-retest reliability for students who first take the test at ages 8–10 is .78 (Schrank, Mather, & Woodcock, 2004). The publisher's directions for individual test administration were adapted for administration on a classroom basis by each classroom teacher with the assistance of a research team member.

*Passage Oral Reading Assessment (PORA).* This assessment measured fluency at the passage level, defined in this study as expressive oral reading of expository or narrative text. Furthermore, it provided data for an alternate measure of syntactic processing.

In the PORA, students' oral rendering of an intact passage, either one of the two passages selected as stimulus texts from the GMRT, was recorded with a Sony digital recorder. The passage that each student was asked to read consisted of the same words as those on the list that the student was given for the WRA. The directions for the PORA were the following:

I'm going to give you a passage to read out loud. Read it as expressively as you can. It's important to make it sound interesting. You don't have to read it quickly. If you come to a word that you don't know, skip it and go to the next word. Continue reading until I ask you to stop or until you finish the passage.

An oral reading fluency rubric was developed to evaluate each student's passage reading on five dimensions. Students were rated on each dimension on a scale ranging from 1 (*very weak*) to 4 (*very strong*). Only two of these dimensions were used in the analyses in the present study: passage expressiveness, which served as our measure of passage-level fluency, and phrasing, which served as our alternate measure of syntactic-level fluency. The other three dimensions evaluated included pace, smoothness, and word expressiveness.

On the passage expressiveness dimension, students' scores were based on their oral interpretation of the passage as a whole, including the appropriateness and consistency of the mood or tone created by their oral reading. If their reading evoked no mood or tone, they received a 1. If approximately a quarter of the passage was interpreted expressively, they received a 2. If half to three quarters of the passage was read with a consistent tone, they received a 3, and if they read the whole or nearly the whole passage in an expressive manner that created a mood or tone that seemed in accord with the author's intention, they received a 4.

The scale for phrasing was drawn largely from the National Assessment of Educational Progress fluency rubric (Pinnell et al., 1995). On this dimension, students receive a 1 if they read primarily word by word, a 2 if they read primarily in two-word phrases, a 3 if they read primarily in three- or four-word phrases or in run-on sentences, or a 4 if they read primarily in larger, meaningful units.

Three judges received training in using the fluency coding rubric and then independently rated the readings of 16 students on the five dimensions. Median agreement for exact scores was 50%, and median agreement for adjacent scores was 96%. A median correlation of .70 was obtained for the ratings of the three judges. To further examine interrater reliability, all ratings of 4 on the original scale were recoded as 3s. On this collapsed scale, exact agreement for two judges (who rated the remainder of the recordings) was 79%.

In sum, the reliabilities of the measures were generally moderate. For background knowledge, word recognition speed, and WJ-III syntactic processing fluency, the test-retest reliabilities ranged from .71 to .79. The test-retest reliability of the GMRT reading comprehension is known to exceed .90. Relatively low reliability was observed for the measure of inferencing, with the alpha of .52 showing error in this measure. Inferencing correlated with GMRT reading comprehension at .57, whereas word recognition speed correlated with it at .65, indicating that inferencing had association with other variables. Although its lower reliability placed limits on the contribution of inferencing to the dependent variable, inferencing retained a statistically significant level of association with comprehension in the full regression model.

## Procedure

At Time 1 (the beginning of the school year), all students in the sample completed the following assessments in the following order: (a) GMRT comprehension subtest, (b) WJ-III Reading Fluency subtest, (c) background knowledge, and (d) inferencing. These assessments were administered by the classroom teacher during regular class time, taking 90 min total across 2 days. In addition, approximately 12 students per classroom completed the WRA and PORA. The students completing these two assessments included all those predesignated as struggling readers (i.e., those assigned to Level 4 of the GMRT because they were deemed to be reading below grade level) and a random sample of at- and above-grade readers (i.e., those assigned to Level 5 or 6 of the GMRT) so that a total of 12 students per class were preselected (however, because of absences, sometimes fewer than 12 students per class were actually tested).

The sample who completed the additional fluency measures had a mean reading comprehension grade equivalent of 5.34 ( $SD =$

Table 2  
*Intercorrelations Between Study Variables*

Variable	1	2	3	4	5	6	7	8	9
Time 1									
1. Word recognition speed	—								
2. Syntactic processing	.62	—							
3. Phrasing	.65	.62	—						
4. Passage-level processing	.57	.55	.69	—					
5. Background knowledge	.50	.55	.52	.51	—				
6. Inferencing	.35	.49	.49	.48	.34	—			
7. Reading comprehension	.65	.75	.68	.67	.67	.57	—		
Time 2									
8. Syntactic processing	.69	.90	.66	.55	.55	.45	.74	—	
9. Reading comprehension	.66	.72	.70	.65	.67	.57	.90	.71	—

Note.  $p < .001$  for all correlations.

3.13). This differed from the total original sample, which had a mean reading comprehension grade equivalent of 5.87 ( $SD = 3.17$ ). Thus, the sample generalizes to students in Grade 5 who are slightly below the Grade 5 students in those schools. The rationale for the selection was that the resources were not available to test all 278 students individually in oral reading fluency within the time allowed for testing by the school. In addition, there was a need to represent the lower achieving readers fully to allow relatively strong inferences to be made about variables that distinguish lower and higher comprehenders in Grade 5.

Research assistants individually administered the WRA and PORA outside of the classroom. Within the three test levels (4, 5, and 6), the order of testing (WRA or PORA first) and the passage presented for the PORA (the expository or narrative passage) were counterbalanced, along with the presentation of the corresponding word list (A or B) for the passage each student was asked to read aloud. The individual testing sessions typically took less than 5 min.

At Time 2, students completed the GMRT comprehension test. The WJ-III Reading Fluency Test was also administered at this time.

## Results

### Variable Information

The following variables from Time 1 were used in the analyses: word recognition speed, syntactic processing, phrasing, passage-level processing, background knowledge, inferencing, and reading comprehension. In the last analysis, syntactic processing and reading comprehension from Time 2 were also used. All variables had distributions within acceptable ranges of normality.

### Descriptive and Correlational Statistics

Table 2 presents the pairwise correlations among all variables included in the analyses. As seen in this table, reading comprehension at Time 1 correlated moderately to strongly with all other Time 1 variables (median  $r = .67$ ), indicating that further analyses of the relationships among these variables were warranted. Specifically, hierarchical regression analyses were planned to partition the shared variance among the variables. In addition, syntactic processing at Time 1 strongly correlated with reading comprehension

at Time 2, as did reading comprehension at Time 1 with syntactic processing at Time 2, indicating that the relationships among these variables might also be explored through hierarchical regression analyses. Table 3 shows the mean, standard deviation, and sample size associated with each variable.

### Question 1

Our first analysis investigated the extent to which each type of fluency—word, syntactic, and passage—correlated with reading comprehension when the other types of fluency and background knowledge and inferencing were statistically controlled. Reading comprehension at Time 1 was entered as the dependent variable in a hierarchical regression with the following independent variables from Time 1 entered as separate steps in this order: background knowledge, inferencing, word recognition speed, syntactic processing, and passage-level processing. As seen in Table 4, which summarizes the statistical output of the five models obtained, the addition of each variable added significantly to the amount of variance accounted for in reading comprehension. Background knowledge alone explained 50% of the variance in reading comprehension, with inferencing accounting for an additional 8%. Word reading speed added another 10% to the explained variance; syntactic processing added 5%; and last, passage-level processing added 2%. Thus, altogether, the five independent variables accounted for 75% of the variance in reading comprehension. Fur-

Table 3  
*Descriptive Statistics for All Variables*

Variable	<i>N</i>	<i>M</i>	<i>SD</i>
Time 1			
Word recognition speed	134	80.37	26.00
Syntactic processing	271	102.53	14.11
Phrasing	145	3.20	0.87
Passage-level processing	145	2.46	1.12
Background knowledge	271	7.91	3.14
Inferencing	272	8.68	2.04
Reading comprehension	270	499.99	49.28
Time 2			
Syntactic processing	271	106.70	17.28
Reading comprehension	278	505.74	51.12



Table 4  
Summary of Hierarchical Regression Analysis for Fluency Variables Predicting Reading Comprehension With Background Knowledge and Inferencing Statistically Controlled at Time 1

Analysis		Final $\beta$ s					Summary statistics with reading comprehension as dependent variable				
Model	Independent variables	BK	INF	WRS	SP	PLP	R	R <sup>2</sup>	$\Delta R^2$	$\Delta F$	dfs
1	BK	.71***	—	—	—	—	.71	.50	.50	126.26***	1, 127
2	BK + INF	.59***	.31***	—	—	—	.76	.58	.08	24.93***	1, 126
3	BK + INF + WRS	.45***	.24***	.36***	—	—	.82	.68	.10	37.25***	1, 125
4	BK + INF + WRS + SP	.36***	.19**	.22***	.32***	—	.85	.73	.05	24.21***	1, 124
5	BK + INF + WRS + SP + PLP	.33***	.15**	.16*	.30***	.19**	.86	.75	.02	10.38**	1, 123

Note. BK = background knowledge; INF = inferencing; WRS = word reading speed; SP = syntactic processing; PLP = passage-level processing.  
\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

thermore, each variable significantly predicted reading comprehension in the final model.

In addition, as seen in Table 4, the betas associated with background knowledge decreased as the number of variables included in the regression models increased. For example, the beta associated with background knowledge in Model 2 was .59 and decreased to .45 when word reading speed was added in Model 3, suggesting that background knowledge was partly associated with word reading speed. Similarly, the betas associated with inferencing, word reading speed, and syntactic processing decreased but remained significant in each subsequent model, providing further indication of both the relatedness and the distinctiveness of the variables.

Question 2

We conducted a second hierarchical regression analysis to analyze the extent to which the association of fluency variables with reading comprehension was mediated by the cognitive variables of inferencing and background knowledge. The dependent variable was reading comprehension, and the independent variables were entered in the following order: word recognition speed, inferencing, background knowledge, syntactic processing, and passage-level processing. The statistical output contained five models, summarized in Table 5.

The first of the five models showed that word recognition speed and reading comprehension were strongly related. The second

model showed that word recognition speed was associated with reading comprehension and inferencing was associated with reading comprehension when each was controlled for the other in the equation. The decrease in association of word recognition speed and reading comprehension from Model 1 to Model 2 indicated that word recognition speed shared variance with inferencing. Note also that word recognition speed retained a unique association with reading comprehension of .54, even after accounting for its association with inferencing.

The third model showed that word recognition speed was also associated with background knowledge. In Model 3, the association of word recognition speed with reading comprehension decreased, which suggested that the effects of word recognition speed on reading comprehension were additionally mediated by background knowledge. Notice that for Model 3, background knowledge was significantly associated with reading comprehension, with its final beta of .45.

The fourth model indicated that syntactic processing was significantly associated with reading comprehension. A reduction in the association of inferencing and background knowledge with reading comprehension from Model 3 to Model 4 was evident. However, all of the contributions of syntactic processing were not mediated because this variable's final beta retained statistical significance.

The fifth model showed that passage-level processing in oral reading shared variance with all of the other variables. Each of the

Table 5  
Summary of Hierarchical Regression Analysis for Assessing Contributions of Fluency Variables to Reading Comprehension and Their Mediated Effects Through Inferencing and Background Knowledge at Time 1

Analysis		Final $\beta$ s					Summary statistics with reading comprehension as dependent variable				
Model	Independent variables	WRS	INF	BK	SP	PLP	R	R <sup>2</sup>	$\Delta R^2$	$\Delta F$	dfs
1	WRS	.66***	—	—	—	—	.66	.44	.44	98.44***	1, 127
2	WRS + INF	.54***	.33***	—	—	—	.73	.53	.10	26.46***	1, 126
3	WRS + INF + BK	.36***	.24***	.45***	—	—	.82	.68	.14	55.45***	1, 125
4	WRS + INF + BK + SP	.22***	.19**	.35***	.32***	—	.85	.73	.05	24.21***	1, 124
5	WRS + INF + BK + SP + PLP	.16*	.15**	.33***	.30***	.19**	.87	.75	.02	10.38**	1, 123

Note. WRS = word reading speed; INF = inferencing; BK = background knowledge; SP = syntactic processing; PLP = passage-level processing.  
\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

variables, including word recognition speed, inferencing, background knowledge, and syntactic processing decreased in strength from Model 4 to Model 5. Yet none of these variables was reduced to statistical insignificance, and passage-level processing retained a statistically significant final beta. Note that for Model 5,  $R = .87$  and  $R^2 = .75$  ( $p < .01$ ).

### Question 3

The next hierarchical regression analysis examined the extent to which each type of fluency—word, syntactic, and passage—correlated with reading comprehension when the other types of fluency were statistically controlled. This analysis served partially as a check on the first analysis in that a different variable (phrasing) was used as a measure of fluency at the syntactic level. Reading comprehension at Time 1 was again entered as the dependent variable, with Time 1 word reading speed, phrasing, and passage-level processing entered in subsequent, separate steps. As seen in Table 6, word reading speed alone explained 43% of the variance in reading comprehension. Phrasing explained an additional 10% of the variance, and passage-level processing explained an additional 4% of the variance beyond that. In the final model, each fluency variable was significantly associated with comprehension.

### Question 4

Finally, hierarchical regression analyses were conducted to address the extent to which fluency predicted change in comprehension over a 12-week period and the extent to which reading comprehension was associated with change in fluency over the same length of time.

In the first analysis, Time 2 reading comprehension was the dependent variable, Time 1 reading comprehension was entered as the first independent variable as an autoregressor, and Time 1 syntactic processing was entered in the second step as a measure of fluency. The output, summarized in Table 7, indicated that Time 1 syntactic processing significantly explained Time 2 reading comprehension, controlling for Time 1 reading comprehension. In the second analysis, Time 2 syntactic processing was the dependent variable, Time 1 syntactic processing was entered as the first independent variable as an autoregressor, and Time 1 reading comprehension was entered in the second step of the regression. The output, which is also shown in Table 7, indicated that Time 1

reading comprehension explained significant variance in Time 2 syntactic processing, controlling for Time 1 syntactic processing.

The use of the autoregressor in these analyses permits the inference that the predictor of syntactic fluency was associated with changes (in this case, increases) in reading comprehension. This is more suggestive of causality than a regression with no autoregressor, in which case the inference is that the predictor was associated with the level of posttest reading comprehension (De Jong & van der Leij, 2002; Gollob & Reichardt, 1987). The same argument holds for the reversed order of testing. That is, with the autoregressor present, pretest comprehension was associated with changes in the syntactic fluency measure. The reciprocity of fluency and comprehension was evaluated for the syntactic measure only as a result of limitations in space available to report the findings. The rationale for using the syntactic measure was that it is a commercially available test and its internal consistency reliability was relatively high.

### Discussion

One of the major findings of the present study was that each of the three types of fluency—at the word, syntactic, and passage levels—related individually to performance on a standardized reading comprehension test in a sample of fifth graders heterogeneous in general reading ability. In other words, the students who demonstrated the highest performances in reading comprehension also displayed (a) fast recognition of isolated words; (b) adeptness in processing phrases and sentences as syntactic units while engaged in oral and silent reading; and (c) appropriate, consistent expression when reading stories and information text out loud. Notably, the relationships between each type of fluency and reading comprehension were significant, both when only the fluency variables were used as predictors in the model and when background knowledge and inferencing skill were included as controls.

Previous studies have reported the association of fluency and comprehension using one measure of fluency, which is usually word recognition speed (De Jong & van der Leij, 2002; Schwanenflugel et al., 2004). However, there remains unexplained variance in reading comprehension in those studies. The present findings suggest that some of the unexplained variance in reading comprehension may be attributed to the two additional forms of fluency that were observed (syntactic fluency and passage level).

The associations between the reading fluency variables and reading comprehension were partially mediated by the cognitive

Table 6  
Summary of Hierarchical Regression Analysis for Fluency Variables Predicting Reading Comprehension at Time 1

Model	Analysis Independent variables	Final $\beta$ s			Summary statistics with reading comprehension as dependent variable				
		WRS	PH	PLP	$R$	$R^2$	$\Delta R^2$	$\Delta F$	$dfs$
1	WRS	.65***	—	—	.65	.43	.43	95.86***	1, 128
2	WRS + PH	.42***	.39***	—	.72	.52	.10	25.45***	1, 127
3	WRS + PH + PLP	.34***	.27***	.26***	.75	.56	.04	11.53**	1, 126

Note. WRS = word reading speed; PH = phrasing; PLP = passage-level processing.

\*\*  $p < .01$ . \*\*\*  $p < .001$ .



Table 7  
*Hierarchical Regression Analyses Assessing Contributions of Fluency to Reading Comprehension Growth and Reading Comprehension to Fluency Growth Using Variables From Time 1 and Time 2*

Analysis 1 (dependent variable = RC-2)								
Model	Independent variables	Final $\beta$ s			Summary statistics			
		RC-1	SP-1	$R$	$R^2$	$\Delta R^2$	$\Delta F$	$dfs$
1	RC-1	.90***	—	.897	.805	.805	1,037.91***	1, 252
2	RC-1 + SP-1	.82***	.10*	.899	.809	.004	5.27*	1, 251
Analysis 2 (dependent variable = SP-2)								
Model	Independent variables	SP-1	RC-2	$R$	$R^2$	$\Delta R^2$	$\Delta F$	$dfs$
1	SP-1	.90***	—	.904	.818	.818	1,107.24***	1, 247
2	SP-1 + RC-1	.82***	.12**	.907	.823	.006	8.21**	1, 246

Note: RC-1 = reading comprehension at Time 1; SP-1 = syntactic processing at Time 1; RC-2 = reading comprehension at Time 2; SP-2 = syntactic processing at Time 2.  $R$ ,  $R^2$ , and  $\Delta R^2$  are specified to three decimal places to illustrate model differences.  
\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

variables of inferencing and background knowledge. The results may be viewed as evidence of an automaticity effect for word recognition speed, as well as an automaticity-like effect for syntactic processing; that is, faster speeds of word recognition and higher performance on the syntactic processing measure (the WJ-III measure of fluency, which is speeded) may indicate that fewer cognitive resources are needed for those activities. Hence, more cognitive resources are available for inferencing and using background knowledge in reading comprehension. The results also indicated that better performance on the passage-level processing fluency measure was linked with better performance on the cognitive measures and two other fluency measures. These links suggest that passage-level fluency may be associated with a greater allocation of cognitive resources to comprehension, but also that passage-level processing is dependent on language processes similar to those involved in inferencing and using background knowledge to understand text.

Last, reading fluency at the beginning of the study predicted growth in comprehension 12 weeks later. In addition, comprehension at the beginning of the study predicted growth in fluency 12 weeks later. In other words, reading comprehension and reading fluency appeared to have a bidirectional relationship when fluency was measured at the syntactic level. Several investigators have recommended that such longitudinal data with the inclusion of an autoregressor are strongly indicative of causal relations (De Jong & van der Leij, 2002; Gollob & Reichardt, 1987); however, previous studies have used students in Grades 1–3 (De Jong & van der Leij, 2002) or have not examined syntactic processing fluency (Jenkins et al., 2003b) with longitudinal data.

The present findings are consistent with automaticity theory's assertion that fast, accurate word recognition frees cognitive resources for reading comprehension (LaBerge & Samuels, 1974). This is suggested by the strong relationships observed in this research between word recognition and reading comprehension performance, the primary type of evidence presented as support by others (e.g., Jenkins et al., 2003b; McCormick & Samuels, 1979).

The findings are also consistent with research suggesting that fluency and comprehension are linked not only because they both involve processing individual words, but also because they both involve processing of syntactic units (Kuhn & Stahl, 2003; Young & Bowers, 1995). We found that two measures of syntactic processing, one involving fast, accurate processing of simple sentences and the other assessing whether students read aloud with proper phrase and sentences units, predicted reading comprehension when controlling for word recognition speed. In addition, the present results extend our current understanding of the relationship between fluency at the syntactic level and reading comprehension, as we also found evidence for an automaticity-like effect for syntactic processing. These findings, on the other hand, appear to be inconsistent with Schwanenflugel et al.'s (2004) findings that prosody in young children's oral reading, measured in part as whether the readers marked phrases and sentences with pauses of appropriate length, did not correlate with reading comprehension after word recognition automaticity was controlled. One reason may be that Schwanenflugel et al.'s (2004) measure of prosody was based on reading a passage that was not part of their comprehension measure, whereas in this study the passage that each student read aloud and on which our phrasing measure was based also appeared on the reading comprehension test that the student received. Thus, in this study the identity of text in the oral reading and the comprehension tasks made it likely that the text features used in fluency might also be used in comprehension.

The present study extended previous research on the links between fluency and comprehension by examining fluency at the whole-passage level. Although some holistic fluency rubrics already existed that give attention to overall expressiveness in combination with the speed and accuracy components of fluency (Pinnell et al., 1995; Zutell & Rasinski, 1991), in this research a new rating scale was created solely for assessing the student's expressive interpretation of the passage. To receive the highest score on this scale, students had to create a tone or mood in their oral rendering of the passage through application of proper stress,

pitch changes, pause structure, and rhythm that was consistent with the author's apparent intent, and they had to maintain this expressive interpretation for the entire length of the passage.

Students who scored high on reading comprehension also tended to receive high scores for expressiveness at the passage level, which indicates that similar types of processing were involved in these tasks. More specifically, this relationship may be an indicator that these students were accurately representing the macrostructure features of the passages they were asked to read aloud. Although having awareness of the macrostructure of both stories and information text has been linked to better reading comprehension (Meyer, 1999; Meyer et al., 1989, 2002; Olson & Gee, 1988), such awareness has not previously been investigated in relation to fluency.

The present study does not attempt to provide direct evidence that students who scored higher on passage expressiveness indeed possessed accurate macrostructures of the passages they read aloud. However, awareness of the macrostructure may explain the relationship between passage expressiveness and reading comprehension that was found after controlling for fluency at the word and syntactic levels along with controlling for other cognitive variables. For example, to illustrate how macrostructure awareness could be manifest through oral reading fluency, consider a child who encounters an information text with the problem-solution structure, for example a text that presents destruction of the rainforests as a problem and explores what can be done to address this problem (Meyer et al., 1989). If the child recognizes that the text possesses this structure, he or she may be more inclined to read the solution section of the passage with a persuasive, excited tone. The child may also place strong emphasis on words that signal the solution (e.g., *answer, to solve these problems*; Meyer et al., 1989).

Finally, the results involving the directionality of the relationship between fluency and comprehension support the idea that these two reading skills have a reciprocally predictive relationship (Stecker et al., 1998). Such findings could be interpreted as support either for the notion that fluency facilitates comprehension growth, in line with automaticity theory, and that comprehension facilitates fluency growth through top-down processes (Rumelhart, 1994) or for joint causation of fluency and comprehension by one or more other factors. The joint causation scenario, it should be noted, aligns with the contention of Paris and colleagues (Paris & Paris, 2001; Paris et al., 2005) that variables such as vocabulary knowledge and motivation may account for high correlations between performance on fluency and comprehension measures. These bivariate correlations support the suggestions of Jenkins et al. (2003b) that comprehension and fluency performances may be linked because they involve similar language processes, perhaps syntactic as well as semantic. Thus, fluency and comprehension skills should become more similar over time up to the age of approximately 10–12 years.

Although the preceding discussion focuses on the contributions of the present study, it is also important to recognize the study's potential limitations. One concern may be multicollinearity because the word recognition and oral reading tasks were developed from the same passages. Furthermore, the passage that each student read aloud was one of the two passages that appeared on the inferencing tests that they received and one of the several passages that appeared on the GMRT. However, we deemed it important to link the measures in this way because of how much the expressive

components of fluency and comprehension may vary from text to text because of differences in the text's grammatical complexity or in the reader's interest in and familiarity with the text. In other words, there seems little reason to expect that expressiveness and comprehension based on very dissimilar texts would be highly related. Furthermore, the bivariate correlations among the variables used in the present study (see Table 1) should assuage the multicollinearity concern; for instance, word recognition speed and passage-level processing had a correlation of .57, which is certainly strong but not so high as to indicate that these measures were assessing the same processes.

Two other limitations of the present study should be mentioned. First, the Cronbach's alpha values for both the inferencing and the background knowledge assessments were lower than the desirable value of .70 or higher. Second, although the regression models accounted for large percentages of the variance in comprehension, important variables may have been left out of the design and thus the analyses. Namely, as mentioned above, Paris and colleagues (Paris & Paris, 2001; Paris et al., 2005) have suggested that vocabulary knowledge and motivation might account considerably for the relations between fluency and comprehension.

In conclusion, the findings of the present study support the multidimensionality of fluency and suggest that future research might examine whether the various levels have different relationships to comprehension across different grade levels. For example, it is possible that word-level fluency is associated with comprehension at primary Grades 1–3, whereas passage-level fluency is associated with comprehension at intermediate Grades 4–6. In addition, future research might examine how the multiple levels of fluency and comprehension are related when motivation is also assessed. Last, assessment of fluency at multiple levels may be particularly important in reading intervention research, as word-, syntactic-, and passage-level fluency may be differentially sensitive to alternative interventions.

## References

- Amble, B. R., & Kelly, F. J. (1970). Phrase reading development training with fourth grade students: An experimental and comparative study. *Journal of Reading Behavior*, 2, 85–93.
- Berninger, V. W., Abbott, R. D., Billingsley, F., & Nagy, W. (2001). Processes underlying time and fluency of reading: Efficiency, automaticity, coordination, and morphological awareness. In M. Wolf (Ed.), *Dyslexia, fluency, and the brain* (pp. 383–414). Timonium, MD: York Press.
- Chard, D. J., Vaughn, S., & Tyler, B. (2002). A synthesis of research on effective interventions for building reading fluency with elementary students with learning disabilities. *Journal of Learning Disabilities*, 35, 386–406.
- Cowie, R., Douglas-Cowie, E., & Wichmann, A. (2002). Prosodic characteristics of skilled reading: Fluency and expressiveness in 8–10-year-old readers. *Language and Speech*, 45, 47–82.
- Cromer, W. (1970). The difference model: A new explanation for some reading difficulties. *Journal of Educational Psychology*, 61, 471–483.
- Daane, M. C., Campbell, J. R., Grigg, W. S., Goodman, M. J., & Oranje, A. (2005). *Fourth-grade students reading aloud: NAEP 2002 special study of oral reading* (NCES 2006–469). Washington, DC: U.S. Government Printing Office.
- De Jong, P. F., & van der Leij, A. (2002). Effects of phonological abilities and linguistic comprehension on the development of reading. *Scientific Studies of Reading*, 6, 51–77.



- Dowhower, S. L. (1987). Effects of repeated reading on second-grade transitional readers' fluency and comprehension. *Reading Research Quarterly*, 22, 389-406.
- Dowhower, S. L. (1991). Speaking of prosody: Fluency's unattended bedfellow. *Theory Into Practice*, 30, 165-175.
- Fleisher, L. S., Jenkins, J. R., & Pany, D. (1979). Effects on poor readers' comprehension of training in rapid decoding. *Reading Research Quarterly*, 15, 30-48.
- Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, 5, 239-256.
- Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education*, 9, 20-28.
- Gollob, H. F., & Reichardt, C. S. (1987). Taking account of time lags in causal models. *Child Development*, 58, 80-92.
- Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101, 371-395.
- Grant, E., & Standing, L. (1989). Effects of rapid decoding training on reading speed and comprehension. *Perceptual and Motor Skills*, 69, 515-521.
- Hannon, B., & Daneman, M. (2001). A new tool for measuring and understanding individual differences in the component processes of reading comprehension. *Journal of Educational Psychology*, 93, 103-128.
- Jackson, N. E., & Doellinger, H. L. (2002). Resilient readers? University students who are poor recoders but sometimes good text comprehenders. *Journal of Educational Psychology*, 94, 64-78.
- Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003a). Accuracy and fluency in list and context reading of skilled and RD groups: Absolute and relative performance levels. *Learning Disabilities Research & Practice*, 18, 237-245.
- Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003b). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology*, 95, 719-729.
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95, 163-182.
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.
- Kintsch, W., & Kintsch, E. (2005). Comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 71-92). Mahwah, NJ: Erlbaum.
- Kuhn, M. R., & Stahl, S. A. (2003). Fluency: A review of developmental and remedial practices. *Journal of Educational Psychology*, 95, 3-21.
- LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 62, 293-323.
- MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates-MacGinitie reading tests: Manual for scoring and interpretation*. Itasca, IL: Riverside.
- Mason, J. M., & Kendall, J. R. (1979). Facilitating comprehension through text structure manipulation. *Journal of Educational Research*, 25, 68-76.
- McCormick, C., & Samuels, S. J. (1979). Word recognition by second graders: The unit of perception and interrelationships among accuracy, latency, and comprehension. *Journal of Reading Behavior*, 11, 107-118.
- Meyer, B. J. F. (1999). The importance of text structure in everyday reading. In A. Ram & K. Morrmann (Eds.), *Understanding language understanding: Computational models of reading* (pp. 227-252). Cambridge, MA: MIT Press.
- Meyer, B. J. F., Middlemiss, W., Theodoru, E., Brezinski, K. L., McDougall, J., & Bartlett, B. J. (2002). Effects of structure strategy instruction delivered to fifth-grade children using the Internet with and without the aid of older adult tutors. *Journal of Educational Psychology*, 94, 486-519.
- Meyer, B. J. F., Young, C. J., & Bartlett, B. J. (1989). *Memory improved: Reading and memory enhancement across the life span through strategic text structures*. Hillsdale, NJ: Erlbaum.
- National Reading Panel. (2000). Fluency. In *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for instruction*. Bethesda, MD: National Institutes of Health, National Institute of Child Health & Human Development.
- O'Shea, L. J., & Sindelar, P. T. (1983). The effects of segmenting written discourse on the reading comprehension of low- and high-performance readers. *Reading Research Quarterly*, 18, 458-465.
- Oakhill, J. V., Cain, K., & Bryant, P. E. (2003). The dissociation of word reading and text comprehension: Evidence from component skills. *Language and Cognitive Processes*, 18, 443-468.
- Olson, M. W., & Gee, T. C. (1988). Understanding narratives: A review of story grammar research. *Childhood Education*, 64, 302-306.
- Paris, S. G., Carpenter, R. D., Paris, A. H., & Hamilton, E. E. (2005). Spurious and genuine correlates of children's reading comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 131-160). Mahwah, NJ: Erlbaum.
- Paris, S. G., & Paris, A. H. (2001). Classroom applications of research on self-regulated learning. *Educational Psychologist*, 36, 89-101.
- Perfetti, C. A., (1985). *Reading ability*. New York: Oxford University Press.
- Perfetti, C. A., & Hogaboam, T. (1975). Relationship between single word decoding and reading comprehension skill. *Journal of Educational Psychology*, 67, 461-469.
- Pikulski, J. J., & Chard, D. J. (2005). Fluency: Bridge between decoding and reading comprehension. *Reading Teacher*, 58, 510-519.
- Pinnell, G. S., Pikulski, J. J., Wixson, K. K., Campbell, J. R., Gough, P. B., & Beatty, A. S. (1995). *Listening to children read aloud: Data from NAEP's Integrated Reading Performance Record (IRPR) at grade 4* (NCES 95-726). Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.
- Rasinski, T. V. (1985). *The study of factors involved in reader-text interactions that contribute to fluency in reading*. Unpublished doctoral dissertation, Ohio State University, Columbus.
- Rasinski, T. V., Padak, N. D., McKeon, C. A., Wilfong, L. G., Friedhauer, J. A., & Heim, P. (2005). Is reading fluency a key for successful high school reading? *Journal of Adolescent & Adult Literacy*, 49, 22-27.
- Rumelhart, D. E. (1994). Toward an interactive model of reading. In R. B. Ruddell, M. R. Ruddell, & H. Singer (Eds.), *Theoretical models and processes of reading* (4th ed.). Newark, DE: International Reading Association.
- Samuels, S. J. (1979). The method of repeated readings. *Reading Teacher*, 32, 403-408.
- Schrank, F. A., Mather, N., & Woodcock, R. W. (2004). *Woodcock-Johnson III Diagnostic Reading Battery: Comprehensive manual*. Itasca, IL: Riverside.
- Schreiber, P. A. (1980). On the acquisition of reading fluency. *Journal of Reading Behavior*, 12, 177-186.
- Schwanenflugel, P. J., Hamilton, A. M., Kuhn, M. R., Wisenbaker, J. M., & Stahl, S. A. (2004). Becoming a fluent reader: Reading skill and prosodic features in the oral reading of young readers. *Journal of Educational Psychology*, 1, 119-129.
- Spring, C., Blunden, D., & Gatheral, M. (1981). Effect on reading comprehension of training to automaticity in word-reading. *Perceptual and Motor Skills*, 53, 779-786.
- Stecker, S. K., Roser, N. L., & Martinez, M. G. (1998). Understanding of oral reading fluency. In T. Shanahan & F. V. Rodriguez-Brown (Eds.),

- 47th yearbook of the National Reading Conference (pp. 295–310). Chicago: National Reading Conference.
- Stein, N. L. (1979). How children understand stories: A developmental analysis. In L. Katz (Ed.), *Current topics in early childhood education* (Vol. 2, pp. 261–290). Norwood, NJ: Ablex.
- Stein, N. L., & Glenn, C. G. (1979). An analysis of story comprehension in elementary school children. In R. O. Freedle (Ed.), *New directions in discourse processing: Advances in discourse processes* (Vol. 2, pp. 53–120). Norwood, NJ: Ablex.
- Thurlow, R., & van den Broek, P. (1997). Automaticity and inference generation during reading comprehension. *Reading and Writing Quarterly*, 13, 165–181.
- Walczyk, J. J., Marsiglia, C. S., Johns, A. K., & Bryan, K. S. (2004). Children's compensations for poorly automated reading skills. *Discourse Processes*, 37, 47–66.
- Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading*, 5, 211–239.
- Young, A. R., & Bowers, P. G. (1995). Individual differences and text difficulty determinants of reading fluency and expressiveness. *Journal of Experimental Child Psychology*, 60, 428–454.
- Young, A. R., Bowers, P. G., & MacKinnon, G. E. (1996). Effects of prosodic modeling and repeated reading on poor readers' fluency and comprehension. *Applied Psycholinguistics*, 17, 59–84.
- Yovanoff, P., Duesbery, L., Alonzo, J., & Tindal, G. (2005). Grade-level invariance of a theoretical causal structure predicting reading comprehension with vocabulary and oral reading fluency. *Educational Measurement: Issues and Practice*, 24, 4–12.
- Zutell, J., & Rasinski, T. V. (1991). Training teachers to attend to their students' oral reading fluency. *Theory Into Practice*, 30, 211–217.

Received September 26, 2006

Revision received July 12, 2007

Accepted August 22, 2007 ■

### Members of Underrepresented Groups: Reviewers for Journal Manuscripts Wanted

If you are interested in reviewing manuscripts for APA journals, the APA Publications and Communications Board would like to invite your participation. Manuscript reviewers are vital to the publications process. As a reviewer, you would gain valuable experience in publishing. The P&C Board is particularly interested in encouraging members of underrepresented groups to participate more in this process.

If you are interested in reviewing manuscripts, please write to the address below. Please note the following important points:

- To be selected as a reviewer, you must have published articles in peer-reviewed journals. The experience of publishing provides a reviewer with the basis for preparing a thorough, objective review.
- To be selected, it is critical to be a regular reader of the five to six empirical journals that are most central to the area or journal for which you would like to review. Current knowledge of recently published research provides a reviewer with the knowledge base to evaluate a new submission within the context of existing research.
- To select the appropriate reviewers for each manuscript, the editor needs detailed information. Please include with your letter your vita. In the letter, please identify which APA journal(s) you are interested in, and describe your area of expertise. Be as specific as possible. For example, "social psychology" is not sufficient—you would need to specify "social cognition" or "attitude change" as well.
- Reviewing a manuscript takes time (1–4 hours per manuscript reviewed). If you are selected to review a manuscript, be prepared to invest the necessary time to evaluate the manuscript thoroughly.

Write to Journals Office, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.



# Native Language Proficiency, English Literacy, Academic Achievement, and Occupational Attainment in Limited-English-Proficient Students: A Latent Growth Modeling Perspective

R. Sergio Guglielmi  
Lake Forest College

The hypothesis that native language (L1) proficiency promotes English acquisition and overall academic achievement, a key theoretical assumption underlying bilingual education, was tested using latent growth modeling of data from 899 limited-English-proficient (LEP) eighth graders who were followed for 12 years in the National Education Longitudinal Study (NELS:88/2000). A model in which L1 proficiency predicted English (L2) reading ability, which in turn predicted high school achievement and distal educational/occupational attainment, fit the data well for the full LEP sample and a Hispanic subsample. In Hispanics, the model explained 24.1%, 7.4%, 29.4%, and 46.3% of the variance in initial English reading level, English reading growth, high school achievement, and post-high school attainment, respectively. Model fit for an Asian subsample, however, was poor. Possible reasons for lack of group invariance include cultural differences in construct conceptualization, greater linguistic and cultural heterogeneity within the Asian subgroup, and cross-language transfer difficulties when L1 and L2 lack a shared alphabetic structure. At least for Hispanic LEP students, this study's results establish the theoretical foundation for exploring the effectiveness of specific educational interventions.

**Keywords:** bilingual education, English language learner, latent growth modeling, multitrait-multimethod, literacy

The U.S. Department of Education (Kindler, 2002) estimates that in 2000–2001 more than 4.5 million public school students (preK–12) in the United States were identified as limited-English-proficient (LEP),<sup>1</sup> and by the year 2030, this number is projected to grow to 40% of the school-age population (Thomas & Collier, 2002). How our schools should respond to the needs of the rapidly expanding LEP population has been the object of a vigorous national debate.

In *Lau v. Nichols* (1974), the U.S. Supreme Court recognized that LEP students would be locked out of the educational system unless schools developed instructional programs that would give them access to a meaningful education despite the language barrier. Although the Court prescribed action, it gave local school

districts the power to decide which particular educational programs to institute. Since then, legislators, educators, academicians, policymakers, and advocacy groups have argued the relative merits of various approaches to educating the growing LEP population. The controversy over the effectiveness of bilingual education has reached especially harsh tones. Fundamentally, the debate centers on the value of two different pedagogical models. One approach is full English immersion. In these cases, the language barrier problem is often addressed by placing LEP students in English-as-a-second-language (ESL) classes, where they receive instruction in the English language skills necessary to operate in a mainstream classroom. On the other side of the debate, proponents of bilingual education advocate academic instruction in both the students' first language (L1) and in English, the second language (L2), with the amount of time spent in L1 instruction decreasing progressively over the course of a few years (early exit programs) or several years (late exit programs).

In the last several years, bilingual education has come under attack, and in some cases (e.g., the passage of Proposition 227 in California and of Proposition 203 in Arizona), its continued existence has been threatened. Unfortunately, judgments about the effectiveness of bilingual education (for or against) are often driven more by sociopolitical motivations than by an objective evaluation of the scientific evidence. A substantial literature on the

---

*Editor's Note.* Linda Baker served as the action editor for this article.—KRH

---

This work was made possible, in part, by a grant from Carnegie Corporation of New York. The statements made and views expressed in this article are solely the responsibility of the author. I am grateful to Lake Forest College for a sabbatical leave that provided the time necessary to complete this project; Nancy Brekke, Michael Nettles, and Sean Reardon for their careful reading of a draft and their constructive feedback; Laura Stapleton for useful suggestions about the analysis of multilevel longitudinal data in a complex sample framework; and Amber Trujillo for her insightful perspective on the challenges and controversies concerning the education of language minorities.

Correspondence concerning this article should be addressed to Sergio Guglielmi, Department of Psychology, Hotchkiss Hall, Box E4, Lake Forest College, 555 North Sheridan Road, Lake Forest, IL 60045. E-mail: guglielmi@lakeforest.edu

---

<sup>1</sup> Although the term English language learner (ELL) has become preferable in the last few years, I will continue to refer to these students as LEP in order to maintain consistency with the terminology most often adopted in legislative documents, in government publications, and in the data set used in the present investigation.

relative value of bilingual versus English immersion instructional approaches is indeed available and has been periodically reviewed (e.g., Rossell & Baker, 1996; Slavin & Cheung, 2003). The methodological difficulties that plague the empirical research, together with the fact that the available evidence does not point incontrovertibly in one direction, however, have left considerable room for political agendas to shape the debate.

Generally speaking, the quality of the research on either side of the controversy is unimpressive. A tolerance, if not predilection, for descriptive and qualitative analyses and for methodological designs that disregard elementary principles of scientific inquiry, has generated findings that are frequently disseminated through outlets that elude the peer review process.

Among other serious weaknesses, it is often difficult to determine the nature of the educational programs to which students were actually exposed. School programs for LEP students may be called bilingual by parents, teachers, or administrators simply because they are received by language-minority students, even when in actuality they are English-only programs, such as structured English immersion or ESL pullout (Rossell, 2003). Moreover, even when a program does indeed have a native language instruction component, calling it bilingual provides little information about its key characteristics. The bilingual programs actually implemented vary tremendously, ranging from minimal L1 instruction and early exit to extensive L1 instruction and late exit (August & Hakuta, 1997; Slavin & Cheung, 2003; Thomas & Collier, 2002).

One way out of the political and scholarly impasse is to shift the research focus away from programs labels and onto specific and definable program components in an effort to identify effective instructional methods, regardless of whether those methods are embedded in bilingual or in English immersion programs (August & Hakuta, 1997, 1998; Slavin & Cheung, 2003). A related approach is to test specific predictions generated from bilingual education theory; without confirmation of such predictions, the theory's validity cannot be established (Cummins, 1999). An important advantage of testing specific hypotheses and program components is that once the key elements of successful outcomes are identified and once the causal structure that specifies the relation among those elements is determined, effective educational programs can be constructed more efficiently and, hopefully, less controversially.

Bilingual education theory allows the prediction that proficiency in one's native language (however achieved) contributes to the development of L2 literacy and ultimately to satisfactory academic achievement in LEP students, regardless of the particular instructional program in which they may have been placed. Some indirect support for this proposition comes from cross-language transfer research.

Additive bilingualism (i.e., the development of fluency in both L1 and L2) has been associated with a variety of general cognitive advantages in the areas of divergent thinking, nonverbal reasoning, concept formation, metalinguistic awareness, creativity, and cognitive flexibility (August & Hakuta, 1998; Portes & Hao, 1998). Positive transfer of literacy skills across languages has also been repeatedly demonstrated (Geva & Siegel, 2000; Holm & Dodd, 1996; Verhoeven, 1994). With respect to the specific cognitive processes involved in biliteracy development, Durgunoğlu (2002), among others, distinguished between (a) language-independent metacognitive and metalinguistic components, which are believed to be universal and, once acquired, support literacy development across different languages and (b) language-specific components,

which reflect instead the particular features of a given language (e.g., its orthographic structure) and may facilitate or hinder the development of literacy in a second language. Phonological awareness is a metalinguistic skill that has been the object of intense research attention, perhaps because of its critical importance in the development of reading. Experimental psycholinguistic research has provided consistent and convincing evidence for cross-language transfer of phonological awareness (e.g., Bialystok, Luk, & Kwan, 2005; Dickinson, McCabe, Clark-Chiarelli, & Wolf, 2004; Durgunoğlu, Nagy, & Hancin-Bhatt, 1993; Gottardo, Yan, Siegel, & Wade-Woolley, 2001).

Although this research indicates that literacy in the first language facilitates reading acquisition in the second language, these conclusions require some qualification. Contrary findings by Gottardo et al. (2001) notwithstanding, several recent studies have reported that phonological awareness and reading skills transfer only, or at least more easily, when L1 shares with English an alphabetic structure (Akamatsu, 2003; Bialystok et al., 2005; Wang & Koda, 2005; Wang, Koda, & Perfetti, 2003). In these studies, participants who spoke an alphabetic first language (i.e., Spanish, Korean, Persian, or Hebrew) exhibited cross-language transfer to English to a much greater degree than did speakers of nonalphabetic first languages (i.e., Chinese or Japanese).

Furthermore, the findings of the literature on cross-linguistic effects come from laboratory research in which participants were exposed to experimental tasks (such as phoneme substitution, phoneme segmentation, semantic category judgment) that arguably bear little resemblance to the tasks faced by LEP students in the classroom. Whether those findings generalize to more complex and more ecologically valid settings and whether the established cross-linguistic transfer effects have a positive and lasting influence on important educational outcomes (e.g., grades, graduation rates) remain open questions. The goal of the present research was to help elucidate these important unanswered issues.

### The Present Study

On the basis of existing research and theory, one can posit a model that specifies the pathways through which bilingual education might exert its effects. According to this model, LEP students' postsecondary academic and occupational attainment is a function of earlier academic achievement, which depends, in large part, on the students' levels of English literacy, particularly as evidenced by reading ability. In turn, L2 literacy is expected to be predicted by proficiency in the native language. To my knowledge, this full model has not been tested in an LEP population, and whatever empirical support exists for individual components of the model comes either from laboratory research or from cross-sectional correlational data. A notable exception is a longitudinal study by Yeung, Marsh, and Suliman (2000) who found that L1 proficiency was unrelated to English scores and to academic achievement. In this study, however, the sample was composed of all students who reported using a language other than English, regardless of whether they were LEP. Moreover, high school and postsecondary outcomes were not included in the model. In the present study, I evaluated the adequacy of the full model by examining longitudinal data from the National Education Longitudinal Study of 1988 (NELS:88/2000) (Curtin, Ingels, Wu, & Heuer, 2002). In particu-



lar, three specific hypotheses, derived from the bilingual education framework, were tested.

1. Self-reported L1 proficiency will predict growth in L2 reading scores from Grade 8 to Grade 12. In view of the relative maturity of the sample, L1 → L2 facilitation effects, if found, would be more likely the result of language-specific mechanisms (e.g., utilization of surface structure similarity between the two languages, transfer of cognitive processing strategies) than the result of universal language structures and nonspecific metalinguistic mechanisms (e.g., cross-language transfer of phonological awareness). Thus, a corollary of the first hypothesis is that L1 proficiency will exhibit stronger L2 literacy facilitation effects in LEP students whose first language shares with English an alphabetic orthography.
2. Self-reported L2 proficiency (which includes students' perceptions of their ability to read in English) will be associated with both eighth grade initial reading scores and rate of reading progress over time.
3. L2 reading ability will predict both high school achievement and long-term educational and occupational end points assessed 8 years after high school graduation. In addition, L2 reading achievement scores will predict post-high school attainment indirectly through the mediation of the student's high school accomplishments.

Tests of these hypotheses required statistical control of socioeconomic status (SES) and of number of grades students completed outside the United States, two covariates expected to influence L1 and L2 proficiency, L2 reading growth, and long-term outcomes. Finally, if the adequacy of the model was confirmed on a full (mixed) sample of LEP students, the empirically established influence of orthographic structure on the extent of cross-linguistic transfer effects required that the model be tested for invariance across language-minority subgroups.

## Method

### *National Education Longitudinal Study (NELS:88/2000)*

NELS:88/2000 is a large-scale longitudinal investigation conducted by the National Center for Education Statistics (NCES). In 1988, more than 25,000 eighth-graders from a multistage, clustered, stratified national probability sample of more than 1,000 public and private schools were administered cognitive tests and questionnaires about their background, school experiences, goals, and attitudes. Supplementary information was obtained from the students' teachers, principals, and parents. The students were then surveyed four more times over a 12-year period (in 1990, 1992, 1994, and 2000), as they progressed through high school (or dropped out) and joined the workforce or went on to pursue postsecondary education. The sample was also "freshened" at each of the first two follow-ups to compensate for attrition and enable both cross-sectional and longitudinal comparisons. In addition, funds made available by the U.S. Office of Bilingual Education and Minority Languages Affairs permitted the oversampling of

Hispanic and Asian/Pacific Islander students. The result is a rich and multifaceted data set that includes thousands of variables and provides information on the lives of nearly 30,000 students, many of whom are not native English speakers. (Full documentation for the NELS:88/2000 project is available at <http://nces.ed.gov/surveys/nels88/>)

### *The LEP Sample*

Proper identification of LEP students from the larger NELS sample was obviously of key importance in this investigation. Although previous research has frequently relied on "bylep"—a composite indicator of base-year LEP status available in the NELS codebook—it is difficult not to share with other investigators dissatisfaction about the performance of this variable. Strang, Winglee, and Stunkard (for Westat, Inc., 1993), for example, looked carefully at the LEP selection issue and concluded that, among other shortcomings, "bylep" yields high rates of false-positive LEP identifications and fails to conform to the legislative definition of LEP as set forth in the Bilingual Education Act of 1988 (Title VII of P.L. 100-297). On the other hand, several elements of Strang et al.'s approach are problematic. Their definition did not incorporate all provisions of the Bilingual Education Act, they considered only language-assistance services received during the eighth grade (and not before) as evidence of a student's difficulty with English, their English proficiency measure cutoff value was not sufficiently stringent, and their definition did not include information from the third NELS wave (since that follow-up had not yet been completed).

Accordingly, an LEP selection measure that mirrors the requirements of the legislative definition was constructed for this study. Figure 1 summarizes the conditions that a student must meet in order to be identified as LEP according to the Bilingual Education Act of 1988.<sup>2</sup> Following a thorough search of the NELS data set (i.e., student, parent, and teacher questionnaires and students' high school transcripts), I extracted 97 variables that, as a group, could be used to assess all of the definitional elements diagrammed in Figure 1. Several of these variables related to the provision that the student's level of English proficiency be insufficient for the student to operate successfully in the classroom. One subset of these variables assessed the students' own perceptions of their ability to understand, speak, read, and write English. A 4-point response scale was provided, ranging from *very well* to *not at all* (although these options differed slightly across waves). Strang et al.'s (1993) LEP definition excluded only the *very well* responses; I excluded both the *very well* and *well* and accepted only the *not very well* and *not at all* responses as positive indicators of LEP status. The inclusion criteria used to select the present LEP sample follow step by step the decision scheme diagrammed in Figure 1. The appli-

<sup>2</sup> Omitted from the summary is the additional provision for "individuals who are American Indian and Alaska Natives and who come from environments where a language other than English has had a significant impact on their level of English language proficiency and who, by reason thereof, have sufficient difficulty speaking, reading, writing, or understanding the English language to deny such individuals the opportunity to learn successfully in classrooms where the language of instruction is English or to participate fully in our society." Such individuals would necessarily meet the other inclusion criteria and would therefore be part of the LEP sample already.

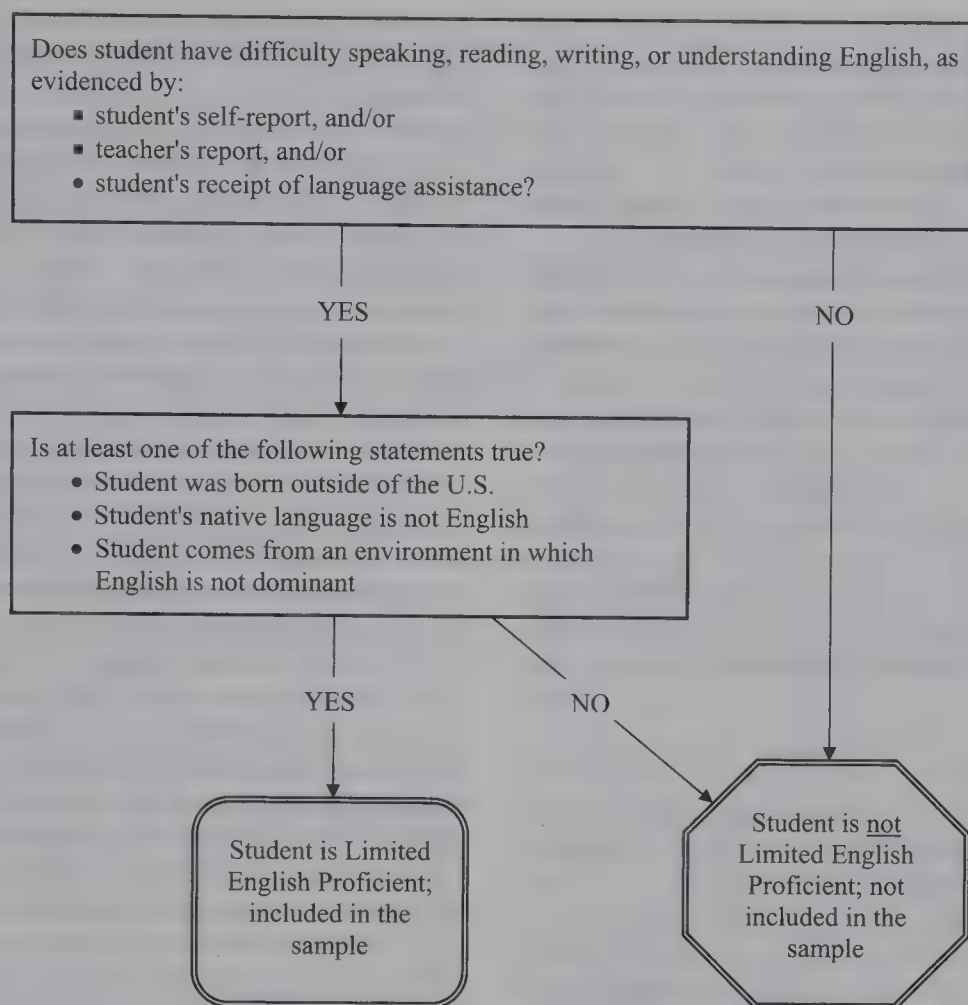


Figure 1. Decision rules for selection of limited-English-proficient sample.

cation of those criteria yielded an LEP sample of 2,997 students. A series of flags was then used to identify LEP responders who participated in all waves of the study and for whom transcript data were available. This process resulted in a final sample of 899 LEP participants.

### Measures

A restricted use data license for the NELS:88/2000 data set from the NCES Office of Data Security permitted access to transcript data and to other individually identifiable information. Many measures used in this research were computed from variables in the NELS data set. Whenever necessary, the original scale was recoded so that high values would reflect high levels of the construct. A conceptual description of these derived indicators is provided below.

### Background Variables

Number of grades completed outside the United States indicates how many grades (0–9), from kindergarten to eighth grade, the student attended outside the United States, according to parental reports. SES is a composite continuous variable available in the NELS data set. It was constructed primarily from base-year parental data (i.e., occupation, educational level, and family income) and is expressed in *z* score units, with higher values indicating higher SES.

### Language Proficiency Measures

Four L1 proficiency and four L2 proficiency measures were collected at Time 1 (Grade 8). For both their first (native) language

and their second language (English), students rated on a 4-point scale their own ability to understand, speak, read, and write. An evaluation of the psychometric properties of these measures is provided in a later section.

### Reading Achievement Test Scores

English reading achievement tests were administered three times (Grades 8, 10, and 12). Although several types of reading test scores can be found in the NELS data files, the most appropriate for longitudinal analyses are the estimated number correct scores generated by item response theory (IRT) procedures. Because their measurement is based on a common scale, these scores are comparable across waves and can be used to gauge achievement gains over time (see Ingels, Scott, Rock, Pollack, & Rasinski, 1994, for details).

### High School Achievement

Four observed variables, measured at Time 3 (Grade 12), were used to assess the relations among individual differences in English reading trajectories and high school completion, rigor of the academic educational program, and overall scholastic achievement.

Number of Carnegie units is a transcript-based indicator that captures, to some extent, the quality of the student's high school curriculum. It expresses the total number of Carnegie units completed by the student in the New Basic subject areas: math, science, English, social studies, computer science, and foreign languages.

Graduation status is a composite measure of high school degree attainment and drop-out history that indicates what type of high



school degree (if any) the student attained and whether the student's degree attainment history was marked by one or more drop-out episodes (0 = *no diploma or equivalent*; 1 = *GED or certificate, with a positive drop-out history*; 2 = *standard high school diploma, with a positive drop-out history*; 3 = *standard high school diploma, with a negative drop-out history*). This measure was based on both self-report and transcript data.

Grade 12 GPA is the student's average grade point average, computed from transcripts, across all math, science, English, and social studies courses taken during the 12th grade. Those academic areas were chosen because they are core subjects, and grades in those areas were available for most students. The restricted-use data set provides standardized grade codes that represent conversions of letter grades and percentages to a common metric. These scores were recoded so that higher values would reflect a higher GPA (1 = *F*, 13 = *A+*).

Class rank is another transcript-based variable that indicates the student's class rank during the last year of school. To account for differences in class size, I converted the original raw scores into percentile ranks.

### *Distal Outcomes*

Three observed measures, collected at Time 4 (i.e., 8 years after graduation from high school), provide an index of the student's long-term post-secondary and occupational attainment.

Postsecondary degree indicates the highest level of post-secondary education attained (0 = *no postsecondary education*; 1 = *some postsecondary education but no degree attained*; 2 = *attained an associate's degree*; 3 = *attained a bachelor's degree*; 4 = *attained a master's degree but not higher*; 5 = *attained a Ph.D. or professional degree*).

An occupational prestige variable was constructed from Time 4 questionnaire data. NELS provides occupational codes and titles for jobs held by participants. The job title list, as well as Appendix D from Nakao and Treas' (1994) occupational prestige study, were given to three undergraduate assistants who were asked to match each job title with the appropriate Nakao-Treas prestige score. I also completed this matching task. Interrater reliability was very high (intraclass correlation = .960). To calculate the final prestige scores, I averaged the four coders' ratings, except in a few cases of obvious coding error when a single aberrant rating was excluded.

Salary is a measure of annual earning rate, computed from two NELS variables: amount earned and pay period (hourly, weekly, biweekly, monthly, or annually).

### *General Analytic Strategy*

When longitudinal data are available, the analysis of change and the modeling of intraindividual and interindividual developmental trajectories are made possible by latent growth curve methods, which operate within the general framework of structural equation modeling (SEM). In this research, latent growth model fitting and estimation procedures were carried out with Mplus Version 4.1 (Muthén & Muthén, 2006). Before the results of these analyses are reported, some critical methodological issues need to be considered.

### *Complex Survey Sample Design*

Because of the multistage, stratified, cluster sample design of the NELS survey, it would be inappropriate to analyze the data as if they

originated from a simple random sample. The biasing effect of unequal probability of selection and nonresponse is one of the analytic challenges presented by this sampling scheme. Sample weighting is necessary to prevent inaccurate estimates of population parameters (Asparouhov, 2005; Stapleton, 2002). In the analyses reported here, the NELS nonresponse-adjusted weight "f4trscwt" was used to weight the cases to population totals. This weight applies to students who participated in the four waves of the study and for whom transcript data were available (Curtin et al., 2002).

Another analytic issue raised by complex multistage sampling designs concerns the multilevel nature of the data they generate. Surveys such as NELS, in which students are nested within schools and schools are nested within larger units (i.e., geographical regions), yield data that have a hierarchical structure. The analytic perils of ignoring the complex data structure have been repeatedly delineated (Hox, 2002; Julian, 2001; Muthén & Satorra, 1995). A measure of the impact that clustering and dependencies among observations have on the precision of variance estimation is provided by the design effect (DEFF). DEFF size is a function of both the intraclass correlation (ICC) and cluster size (Julian, 2001; Muthén & Satorra, 1995). Although in the present study the average cluster size was only 2.31, ICCs ranged from .09 to .28. Some of the resulting DEFFs were thus large enough that failure to account for the hierarchical structure of the data could have produced biased standard error estimates.

Two data analytic strategies are available in Mplus when the data to be modeled have a complex sample structure. The "type = complex" option does not provide estimates of between-level and within-level variance components, but it permits specification of stratum, cluster, and weight variables needed to generate standard errors and goodness-of-fit indices that take into account the nonindependence of observations associated with stratified cluster sampling designs. In a Monte Carlo simulation, Muthén and Satorra (1995) found that this analytic option is particularly robust to violations of nonnormality and is suitable for a variety of sampling designs and types of data (e.g., dichotomous variables). This solution to the complex design problem was preferred to the "two-level" alternative because the key issue examined here was the extent to which L1 proficiency, developed prior to eighth grade, predicts academic and vocational achievement in LEP students; the focus was not on school-level effects, which would require modeling within-school and between-schools variance components in two-level analyses. Moreover, when the type = complex analytic strategy is chosen, Mplus (Version 3.1 and later) allows the option of including a stratification variable, in addition to cluster and sampling weight information. In a series of simulation studies, Asparouhov (2004) demonstrated that failure to include stratum information in the analysis (even when cluster is specified) can markedly influence model rejection rates. Similar conclusions were reached more recently by Stapleton (2006), who conducted a simulation study in which five estimation approaches, in the presence of complex sample data, were compared under different analytic scenarios. She found that the estimation method used in the present research yielded the most robust parameter and standard error estimates under all analytic conditions.

### *Missing Data and Nonnormality*

Invariably, longitudinal research presents the problem of what to do with missing data. Reviews and empirical comparisons of methodological approaches for dealing with missing data (e.g., Enders &



Bandalos, 2001; Newman, 2003; Schafer & Graham, 2002) uniformly point to the same general conclusions. When data are missing completely at random (MCAR) or when data are missing at random (MAR), multiple imputation (MI) and maximum likelihood (ML) estimation procedures greatly outperform traditional approaches, such as pairwise deletion, listwise deletion, mean imputation, and hot deck procedures. Full information maximum likelihood (FIML), a particular ML algorithm implemented in many leading SEM software packages (including Mplus), has been found especially effective (Enders & Bandalos, 2001). Although the MCAR missingness assumption rarely holds in applied research (Kline, 2005), Schafer and Graham (2002) suggested that it is appropriate to proceed under the less restrictive and more realistic MAR assumption. As they conclude in their authoritative review, "ML and MI under the MAR assumption represent the practical state of the art" (p. 173). This was the approach used in the present study.

In addition to MAR data, multivariate normality is a key assumption of ML estimation. Mplus can handle departures from normality in the presence of missing data by computing robust standard errors with a sandwich estimator and applying a scaling correction to adjust the chi-square test statistic (Muthén & Muthén, 2006). This corrected chi-square is the Yuan–Bentler test statistic (Yuan & Bentler, 2000), an extension of the Satorra–Bentler scaled chi-square ( $\chi^2_{SB}$ ) to the case of missing data.<sup>3</sup> Bentler and his associates (Chou & Bentler, 1995; Gold, Bentler, & Kim, 2003; Savalei & Bentler, 2005) found that when the data are incomplete and nonnormal and when the mechanism of missingness is MCAR or MAR, ML estimation with the Yuan–Bentler correction provides accurate parameter estimates, standard errors, and test statistics.

### *Assessment of Model Fit*

The results of covariance structure analyses with large samples cannot be properly evaluated by means of the  $\chi^2$  goodness-of-fit statistic alone. Given a sufficiently large sample size, any model can be rejected on the basis of the  $\chi^2$  test, no matter how small the discrepancy between the observed and the estimated covariance matrices (Kline, 2005). Although many alternative fit indices are now available, and their relative merits continue to be debated (e.g., Fan & Sivo, 2005; Marsh, Hau, & Wen, 2004), the recommendations issued by Hu and Bentler (1998, 1999) have gained wide acceptance. Consistent with those recommendations, the indices used in the present research include the comparative fit index (CFI; Bentler, 1990), the root-mean-square error of approximation (RMSEA; Steiger & Lind, 1980), and the standardized root-mean-square residual (SRMR; Bentler, 1995). With respect to cutoff values for these indices, Hu and Bentler (1999) recommend that CFI be close to .95, RMSEA be close to .06, and SRMR be close to .08.

### *Results*

The two language proficiency constructs are at the core of the theoretical model tested in this study. Thus, before reporting the results of the latent growth analyses, it is important to examine the psychometric characteristics of those self-report measures. Most of the other variables used in this study were derived either from student transcripts or from the cognitive test battery that was administered three times over a 4-year period. The reliability and validity of those instruments have been fully evaluated and are

clearly established (Rock & Pollack, 1991, 1995). Technical reports by the NCES (Kaufman & Rasinski, 1991; McLaughlin & Cohen, 1997) have in many cases also reached favorable conclusions about the psychometric properties of the student questionnaire data. The findings of those evaluations, however, tend to be domain specific, wave specific, and source specific and do not allow general conclusions about the quality of the self-report data, particularly of the language proficiency measures.

A substantial linguistic literature points to the criterion-related validity of language proficiency self-assessments. In a review of 21 studies, Blanche and Merino (1989) found that typical correlations between self-ratings and objective indicators of L2 proficiency range between .50 and .60. The results of a more recent meta-analysis by Ross (1998) were consistent with Blanche and Merino's conclusions and indicated that for each of the four language skills (i.e., understanding, speaking, reading, and writing), validity coefficients range from .51 to .65. Very similar and even higher validity coefficients have been reported for self-assessment of L1 proficiency (Delgado, Guerrero, Goggin, & Ellis, 1999; Shameem, 1998). In a study by Hakuta and D'Andrea (1992) that is particularly relevant to the present research, 308 high school students of Hispanic origin were asked to provide ratings of both their Spanish proficiency and their English proficiency, which were then compared with standardized proficiency measures. The validity coefficients were .61 for L1 and .46 for L2. Although these findings offer general support for the psychometric adequacy of language proficiency self-assessments, the reliability and validity of the NELS-based language proficiency measures used in this research need to be directly evaluated.

### *Psychometric Properties of NELS:88/2000 Language Proficiency Measures*

#### *Reliability*

Students who indicated that a language other than English was spoken at home were asked to rate their ability to understand, speak, read, and write that language as well as English. Self-ratings of L1 proficiency were taken at base year (Time 1) and again 2 years later at the first follow-up (Time 2). English proficiency ratings were taken three times, at intervals of 2 years. Excellent reliability and stability were found for both L1 and L2 proficiency measures at all assessment waves. Internal consistency (Cronbach's alpha) for L1 proficiency was .86 at both Time 1 and Time 2. Alpha coefficients for L2 proficiency were .91, .93, and .90 at Time 1, Time 2, and Time 3, respectively. For each of the four language skills that make up the L1 proficiency construct, the correlations between Time 1 and Time 2 measures ranged from .55 to .70 ( $M = .63$ ). For L2 skills, correlations between adjacent waves yielded stability coefficients ranging from .35 to .56 ( $M = .50$ ), whereas correlations between Time 1 and Time 3 measures ranged from .26 to .54 ( $M = .44$ ). It is plausible that differences in self-ratings over time reflect not just measurement error but also individual differences in language proficiency trajectories, which would then attenuate the correlations, especially between nonadjacent assessment occasions.

<sup>3</sup> Unless otherwise noted, all chi-squares reported in this article are computed with the Yuan–Bentler scaling correction.



Table 1  
*Multitrait–Multimethod Correlation Matrix for Two Assessment Methods and Two Traits Measured at Time 1 and at Time 2*

Method/measure	Self-report				Objective assessment			
	1	2	3	4	5	6	7	8
Self-report								
1. L2 proficiency at T1	—							
2. Non-English grades at T1	-.051	—						
3. L2 proficiency at T2	.542	.029	—					
4. Non-English grades at T2	-.038	.470	-.064	—				
Objective assessment								
5. IRT reading scores at T1	.297	.328	.324	.247	—			
6. Non-English GPA at T1	.085	.379	.134	.499	.413	—		
7. IRT reading scores at T2	.261	.307	.315	.273	.773	.451	—	
8. Non-English GPA at T2	-.026	.379	.082	.578	.311	.611	.392	—

Note. Because of the large sample size ( $N = 1,804$ ), correlations as small as .047 are significant at  $p < .05$ . T1 = Time 1 data collection; T2 = Time 2 data collection; IRT = item response theory, GPA = grade point average.

### Construct Validity: A Multitrait–Multimethod Model

A second important issue concerns the construct validity of the L1 and L2 proficiency measures that are based on self-report and are therefore vulnerable to selective distortion by self-presentational biases and other response sets. Students' self-rated proficiency measures reflect three sources of variance: (a) trait variance, which is systematic variance, independent of assessment method, that represents true differences in ability level; (b) method variance, which is systematic variance specifically associated with the use of self-report (e.g., the tendency to be influenced by social desirability concerns); and (c) error variance, which includes both residual systematic variance and measurement error (e.g., misreading or misunderstanding a question). In order to determine whether self-ratings of language proficiency reflect true variations in ability, one must disentangle the contributions of trait effects, method effects, and error effects. Moreover, if L1 or L2 proficiency were found to predict favorable academic and occupational outcomes, it would be important to rule out the possibility that those associations were simply the byproduct of a general intellectual ability factor.

The multitrait–multimethod (MTMM) approach pioneered by Campbell and Fiske (1959) can be used to decompose the total variance and to assess the convergent and discriminant validity of psychological measures. This design requires that two or more traits be assessed with two or more methods. Regrettably, the NELS data set provides no measure of L1 proficiency other than self-report. More objective indices of proficiency, against which one could validate students' self-ratings, however, are available for L2. Table 1 shows a MTMM matrix that includes two methods (self-report and objective measurement) and two traits (L2 proficiency and achievement in academic subjects other than English) assessed at two points in time (Grades 8 and 10).<sup>4</sup> The objective indicators of the two traits are based on transcripts and on standardized test scores. The non-English achievement indicators, included to test the discriminant validity of the L2 proficiency measure, represent self-reported and transcript-based grades in math, science, and social studies courses.

Longitudinal confirmatory factor analysis (CFA) of the MTMM matrix was used to evaluate the construct validity and temporal stability of self-reported L2 proficiency. The CFA model (see

Figure 2) includes two trait factors across two waves, two method factors, and eight observed variables. For each wave, one self-report variable and one objective assessment variable were specified to load on each trait.<sup>5</sup> The manifest variables that at each wave were hypothesized to load on the L2 proficiency trait factor were the IRT-estimated reading scores and a self-reported proficiency measure computed by averaging students' self-rated ability to understand, speak, read, and write English.<sup>6</sup> The two indicators that were specified to load at each wave on the Non-English Achievement trait factor were (a) non-English grades, which were computed by averaging students' self-reported grades in math, science, and social studies, and (b) non-English GPA, which was

<sup>4</sup> Although L2 proficiency ratings are available also at Time 3 (second follow-up questionnaire), a self-report measure of grades in various academic subjects was not included at that time. Thus, only the first two waves of data could be used to test the model.

<sup>5</sup> There is considerable debate on the advantages and disadvantages of various CFA parameterizations of MTMM matrices (see Brown, 2006, for a recent review). The correlated trait-correlated method (CT-CM) approach was used in the present research for the theoretical and substantive reasons articulated by Conway, Lievens, Scullen, and Lance (2004) and by Lance, Noble, and Scullen (2002). This parameterization, however, is notorious for returning ill-defined solutions (Kenny & Kashy, 1992). The problem is often remedied with the use of large samples (e.g., Lance et al., 2002). Since the data to be modeled no longer required participation in all waves of the study, the CFA–MTMM model was tested on 1,804 participants who met the inclusion criteria described in Figure 1, participated in the first and second follow-ups, and had transcript information. The data were weighted with "f2trp1wt," the appropriate NELS weight in panel analyses in which survey and cognitive test data are combined with transcript data.

<sup>6</sup> In an alternative model, the mean of the four language skills was replaced with only the reading self-rated skill that was specified to load, together with the IRT reading measure, on a L2 reading proficiency trait factor. Estimation of this model produced a "not positive definite matrix" solution, probably as a result of a very high correlation between Time 1 and Time 2 reading proficiency factors ( $r = .993$ ). The fit of this model, the parameter estimates, and the standard errors, however, were substantively interchangeable with those of the model shown in Figure 2.

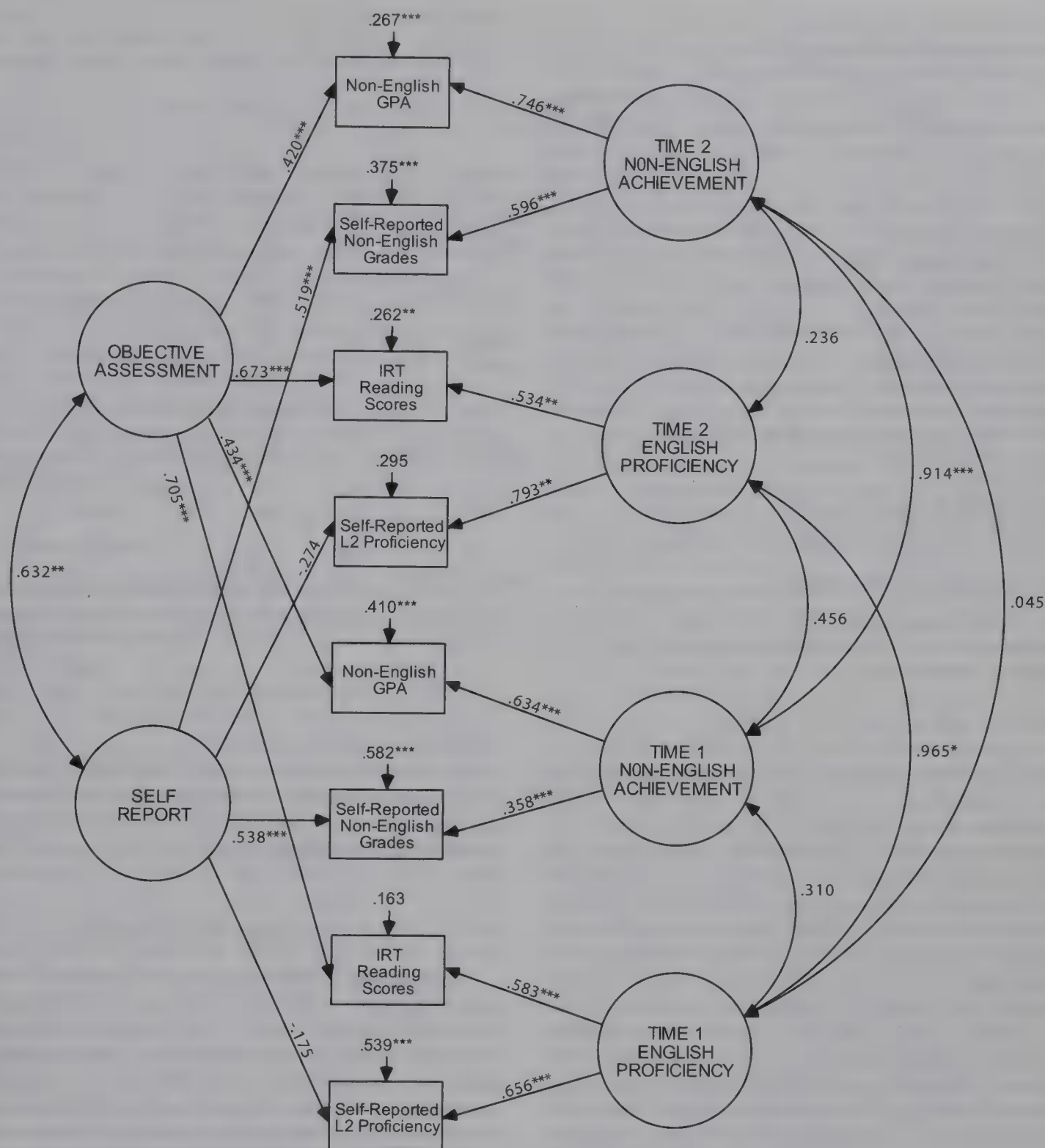


Figure 2. Multitrait-multimethod confirmatory factor analysis model of two correlated traits and two correlated methods across two measurement waves. Completely standardized robust maximum likelihood parameter estimates. The residual variance components (error variances) indicate the amount of unexplained variance. Thus, for each observed variable,  $R^2 = (1 - \text{error variance})$ . GPA = grade point average; IRT = item response theory; L2 = English. \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

the transcript-based average of students' GPA in math, science, and social studies.<sup>7</sup>

The fit of the CFA-MTMM model displayed in Figure 2 was excellent,  $\chi^2(5, N = 1,804) = 10.236$ , CFI = .994, RMSEA = .024, SRMR = .015 (see Model 1 in Table 2, which summarizes the fit of all the models tested in the present investigation). For both waves, all measures loaded significantly ( $p < .001$ ) on their respective trait factors, even after shared method variance was

controlled. This indicated good levels of convergent validity. Discriminant validity of the measures was demonstrated by the non-

<sup>7</sup> The non-English GPA variable for Time 2 is an average of standardized course grades obtained in 10th grade. The NELS transcript file, however, includes eighth-grade GPA for only a handful of students; thus ninth-grade courses were used to derive the Time 1 non-English GPA variable.



Table 2  
Fit Indices for the Models Tested

Model	Comparison model	$\chi^2$	df	CFI	RMSEA	SRMR	$\Delta\chi^2$	$\Delta df$	$\Delta CFI$
Single group analyses									
1. MTMM	—	10.236	5	.994	.024	.015	—	—	—
2. Growth	—	4.290*	1	.997	.061	.015	—	—	—
3. CFA	—	190.213*	82	.961	.038	.043	—	—	—
4. Full conditional LGM	—	343.074*	150	.952	.038	.050	—	—	—
5. Full LGM (Hispanic sample)	—	355.490*	150	.934	.055	.064	—	—	—
6. Model 5, but paths from L1 proficiency to growth factors fixed at 0	5	391.113*	152	.923	.059	.092	79.052*	2	.011
7. Model 5, but paths from L2 proficiency to growth factors fixed at 0	5	361.884*	152	.932	.055	.070	6.248*	2	.002
8. Full LGM (Asian sample)	—	641.974*	150	.778	.114	.098	—	—	—
9. Model 8, but paths from L1 proficiency to growth factors fixed at 0	8	Improper solution (negative $\Delta\chi^2$ )							
10. Model 8, but paths from L2 proficiency to growth factors fixed at 0	8	643.184*	152	.778	.113	.103	5.218	2	.000
Multigroup invariance tests									
11. Unconstrained growth model—baseline	—	2.632	2	.998	.030	.016	—	—	—
12. Model 11, but growth factor means invariant	11	7.027	4	.990	.046	.073	4.018	2	.008
13. Model 12, but growth factor variances & covariances invariant	12	10.538	7	.988	.038	.084	3.529	3	.002
14. Unconstrained CFA model—baseline	—	580.206*	164	.889	.085	.070	—	—	—

Note.  $\chi^2$  = Yuan-Bentler corrected  $\chi^2$ ; CFI = comparative fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root-mean-square residual; MTMM = multitrait-multimethod approach; CFA = confirmatory factor analysis; LGM = latent growth modeling; L1 = native language; L2 = English.  
\* $p < .05$ .

significant correlations between different trait factors, both within and across measurement occasions. At the same time, the very high correlations for the same trait factor across time indicated excellent temporal stability of the constructs. Table 1 shows that heterotrait-monomethod correlations were generally higher in the objective assessment block than in the self-report block, indicating greater shared method variance for the objective measures than for the self-report measures. The squared factor loadings of each measure on its respective trait and method factors made it possible to partition the total variance into trait, method, and error components. Again, large method effects were evident for the objective measures, all of which loaded significantly ( $p < .001$ ) on their method factor. Method effects were smaller but also significant for the self-report non-English grade measures and were nonsignificant for the self-report L2 proficiency indicators. Noteworthy was the large amount of trait variance relative to method variance for both waves, particularly for the self-report measures. Overall, averaging variance components across all measures and across the two measurement occasions indicated that 39.1% of the variance in the measures was explained by trait variance, 24.7% was explained by method variance, and the remaining 36.2% was due to error.

Taken together, these findings are reassuring with regard to the reliability, stability, and construct validity of the L2 proficiency self-ratings. The L2 proficiency measure loaded strongly on the same trait factor on which IRT-estimated English reading scores loaded significantly. In addition, the nonsignificant correlations between L2 proficiency and academic achievement in non-English subjects suggest that L2 skills were not byproducts of a general cognitive ability factor. It seems reasonable to extend to the L1 proficiency measure the favorable conclusions reached about the

psychometric characteristics of the L2 proficiency variable, particularly considering that criterion-related validity coefficients have repeatedly been found to be higher for L1 than for L2 self-assessment (e.g., Delgado et al., 1999; Hakuta & D'Andrea, 1992).

Latent Growth Curve Analyses

The key hypothesis under examination in this study was that L1 proficiency would predict the development of L2 reading skills in LEP students, and this, in turn, would be associated with successful academic and occupational outcomes. IRT English reading scores from three points in time—8th, 10th, and 12th grades—were available. Thus, the first step was to establish whether this sample of LEP students showed evidence of growth in reading scores and whether there were substantial individual differences in growth. Furthermore, prior to evaluating the structural component of the model, the validity of the measurement model needed to be assessed. Assuming confirmation that the manifest indicators adequately measured their respective latent factors and assuming a reasonable level of interindividual variation in initial reading scores and rate of change, one could then evaluate the hypothesized relations among latent constructs.

The longitudinal structure of the data makes these research questions ideally suited for latent growth modeling (LGM) analyses, which become possible when at least three measurement occasions are available for the repeated measure variable. Although multistep modeling strategies have been the object of some controversy (e.g., see *Structural Equation Modeling* [2002], 7, [1], the analytic approach adopted in the present research followed

Tomarken and Waller's (2003) sensible recommendation to parse complex SEM models into smaller and theoretically meaningful components.

### *Descriptive Statistics*

All 899 LEP students were included in the analyses. Means, standard deviations, and intercorrelations for the 20 observed variables are reported in Table 3. Although some variables were not normally distributed, transformations were not applied because, as discussed earlier, the complex sample analytic option in Mplus automatically invokes the "MLR (maximum likelihood robust) estimator," which provides FIML parameter estimates and standard errors that are robust against violations of distributional assumptions (Muthén & Muthén, 2006). The only exception was the salary variable. Its very large variance, skew, and kurtosis values caused convergence problems during model fitting. Dividing salary values by 10,000 and applying a cubic root transformation remedied the problem.

### *Single Group Analyses*

The analytic steps outlined above were first applied to the full LEP sample ( $N = 899$ ). The issue of group differences will be considered in a later section.

*The growth model.* Figure 3 shows a simple two-factor unconditional (i.e., no covariates specified) latent growth model in which reading scores at three points in time are indicators of two latent growth factors, reading intercept and reading slope. The first factor represents the initial level of English reading achievement (i.e., eighth-grade IRT reading scores), and the second factor represents the slope of individual trajectories in the development of English reading ability over time. For any given student, the initial reading value remains the same across the three time points; therefore, as is customary in LGM, the three intercept loadings were fixed to 1. The assumption of linear change in reading scores across time was modeled by fixing the three slope loadings to 0, 1, and 2 for the first, second, and third measurement occasion, respectively.

Estimation of the model displayed in Figure 3 yielded a very good fit,  $\chi^2(1) = 4.290$ , CFI = .997, RMSEA = .061, SRMR = .015 (Model 2, Table 2). Although there was a tendency for initial L2 reading scores to be positively associated with rate of reading improvement ( $r = .214$ ), the estimated covariance between intercept and slope factors was not significant. On the other hand, the means of both growth factors were significantly different from zero. The mean reading score at the first measurement occasion (i.e., the initial value) was 21.01,  $p < .001$ , and the mean rate of improvement over time was 2.57 units for each measurement occasion,  $p = .01$ . In addition, the variances of both growth factors were significantly different from zero ( $p < .001$  and  $p < .01$  for the intercept and slope, respectively). The latter finding indicated the presence of significant interindividual variation in both initial reading level and rate of growth from Grade 8 to Grade 12, and it obviously raised the question of what might account for that variance, an issue that will be fully elaborated in a subsequent sentence.

*The measurement model.* A four-factor CFA model that included all the latent constructs (except for the two growth factors) to be incorporated into the full structural model, together with their

respective indicators, is shown in Figure 4. Time 1 self-ratings of the ability to speak and understand either English or one's native language reflect oracy skills that are likely to share common omitted causes. Thus, for the two language proficiency constructs, the residuals of oracy indicators (speaking and understanding) were allowed to correlate not only because of their large modification indices, but also because such model modification was justified on theoretical grounds (e.g., Delgado et al., 1999). Because of convergence difficulties, however, the uniquenesses of the literacy measures (reading and writing) were not allowed to correlate. The fit of the measurement model diagrammed in Figure 4 was very good,  $\chi^2(82) = 190.213$ , CFI = .961, RMSEA = .038, SRMR = .043 (Model 3, Table 2). Thus, the hypothesized relations of manifest indicators to their underlying factors were confirmed. All factor loadings were significant at  $p < .001$  or better and in the expected direction. For both L1 and L2, reading and writing self-ratings had the highest loadings (all  $> .90$ ) on the proficiency factors, which indicated that literacy components played a very prominent role in the definition of language proficiency. For both L1 and L2 literacy indicators, the proportion of variance explained by their respective latent factor ranged from .827 to .908 (see Figure 4).

Assessment of the measurement model provided good evidence for both convergent and discriminant validity. All of the observed measures of the same construct loaded significantly on that construct, and the low intercorrelations among factors (except for the expected covariation between the two achievement factors) indicated that those factors represent different constructs. Having established the validity of the model's factorial structure, it was then appropriate to examine the structural relations among constructs.

*The full conditional LGM model.* The key assumptions of bilingual education theory served as the theoretical foundation for specification of the full latent growth model presented in Figure 5, which includes both measurement and structural components. Self-reported L1 and L2 proficiency were hypothesized to explain variation in reading growth. L1 proficiency was expected to predict the initial level (L2 reading intercept) and, more important, the developmental trajectories of students' L2 reading achievement (L2 reading slope). L2 proficiency was also hypothesized to have a direct relation with the latent growth factors that, in turn, were expected to predict both high school achievement and more distal educational and occupational sequelae. Direct effects of L1 and L2 proficiency on high school and post-high school achievement were not specified because the language proficiency measures were hypothesized to predict academic and occupational attainment through the development of English reading skills. SES and number of grades completed outside the United States prior to the eighth grade are time-invariant covariates that were added to the model because they might predict individual differences in L1 and L2 proficiency, reading growth factors, and high school achievement. The full LGM model fit the sample data quite well,  $\chi^2(150) = 343.074$ , CFI = .952, RMSEA = .038, SRMR = .050 (Model 4, Table 2). Completely standardized estimates, together with their significance levels, are included in Figure 5.

Self-reported L1 proficiency was significantly predicted by number of grades completed outside the United States, as was L2 proficiency, but the latter path coefficient was negative. These



Table 3  
Descriptive Statistics and Intercorrelations for Observed Variables

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1. No. of grades outside United States	—																			
2. Socioeconomic status	.05	—																		
3. Understand L1	.17	-.14	—																	
4. Speak L1	.19	-.20	.80	—																
5. Read L1	.25	-.21	.52	.60	—															
6. Write L1	.24	-.20	.49	.59	.89	—														
7. Understand L2	-.25	.24	.10	.01	-.02	-.09	—													
8. Speak L2	-.26	.22	.06	-.01	-.14	-.15	.77	—												
9. Read L2	-.35	.13	.04	-.02	-.08	-.12	.71	.77	—											
10. Write L2	-.34	.11	.05	.01	-.03	-.04	.71	.74	.85	—										
11. IRT reading Grade 8	-.04	.23	.10	.09	.10	.08	.34	.18	.24	.23	—									
12. IRT reading Grade 10	-.03	.24	.11	.10	.11	.07	.28	.14	.22	.19	.80	—								
13. IRT reading Grade 12	.06	.21	.17	.16	.18	.14	.27	.13	.17	.15	.76	.84	—							
14. No. of Carnegie units	.13	.35	.04	.01	.02	.01	.10	.01	-.01	.02	.40	.43	.45	—						
15. Graduation status	-.05	.20	-.05	-.01	.05	.02	.13	.01	.06	.10	.25	.23	.23	.71	—					
16. Grade 12 GPA	.02	.26	-.03	-.03	-.08	-.06	.00	-.04	-.02	-.01	.26	.33	.33	.66	.61	—				
17. Class rank	.10	.25	.07	.06	.01	-.01	.06	-.02	.00	.00	.41	.43	.44	.71	.50	.72	—			
18. Postsecondary degree	.09	.33	.02	.03	.02	-.01	.14	.00	.04	.02	.42	.44	.47	.57	.39	.45	.50	—		
19. Occupational prestige	-.01	.20	-.04	-.02	.00	.01	.05	.03	.03	.04	.22	.19	.22	.29	.22	.34	.33	.40	—	
20. Salary	.06	.18	.10	.07	.02	.03	.04	-.02	-.01	-.05	.20	.23	.23	.18	.10	.17	.17	.32	.19	—
<i>M</i>	.46	-.52	3.50	3.29	2.76	2.63	3.53	3.47	3.44	3.41	21.06	23.22	26.22	12.91	2.47	6.24	4.17	1.19	45.76	1.31
<i>SD</i>	1.46	.84	.64	.75	1.05	1.11	.70	.73	.80	.83	7.31	8.75	9.69	5.70	.99	2.80	3.03	1.02	13.21	.26

Note. Because of the large sample size ( $N = 899$ ), correlations as small as .07 are significant at  $p < .05$ . L1 = native language; L2 = English; IRT = item response theory.

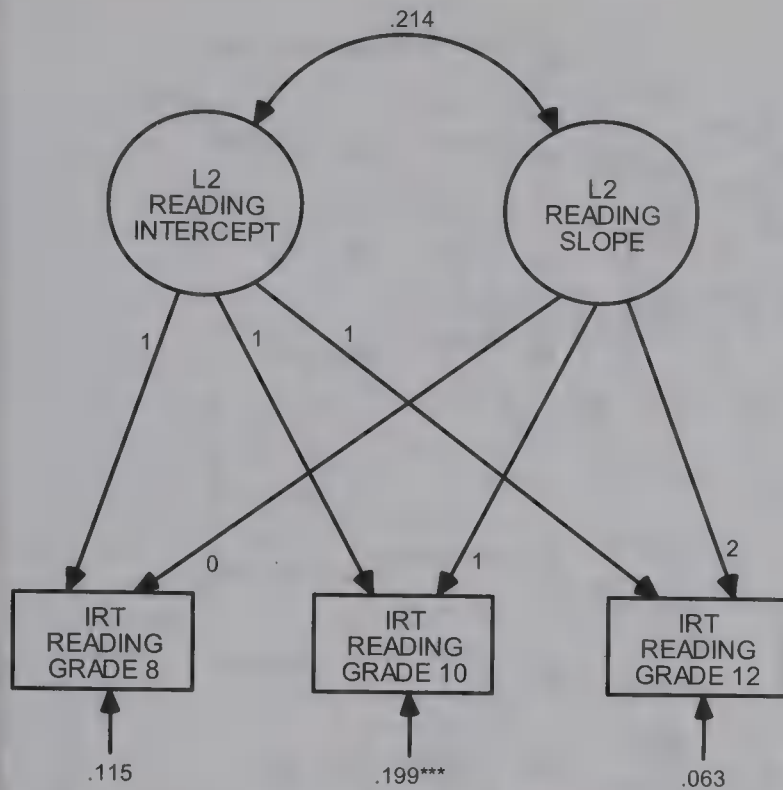


Figure 3. Unconditional two-factor linear growth model. Completely standardized robust maximum likelihood parameter estimates. The residual variance components (error variances) indicate the amount of unexplained variance. Thus, for each observed variable,  $R^2 = (1 - \text{error variance})$ . L2 = English; IRT = item response theory. \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

findings were not unexpected; the longer students had been schooled abroad, the greater their native language proficiency and the lower their English proficiency (or, at least, the lower their confidence in their English proficiency). Paths from SES to both language proficiency measures were also significant, but in the opposite direction (i.e., negative for L1 proficiency and positive for L2 proficiency). From a theoretical perspective, the most interesting findings relate to the test of the bilingual education assumptions. As expected, L2 proficiency was a significant predictor of initial L2 reading level, but it did not predict improvement in reading over time. On the other hand, L1 proficiency significantly predicted both initial English reading status and rate of English reading growth from Grade 8 to Grade 12, even after the influence of being schooled abroad and SES was controlled. Figure 5 also shows significant path coefficients from the two latent growth factors to high school achievement, from high school achievement to post-high school achievement and from both growth factors to post-high school achievement.<sup>8</sup> Overall, the variables in the model explained 12.9%, 3.9%, 30.7%, and 55.9% of the variance in initial English reading scores, English reading rate of improvement, high school achievement, and post-high school achievement, respectively (see Figure 5).

### Multigroup Analyses

The operating assumption, up to this point, has been that the LEP sample examined in this study was a homogeneous group and that the model parameters (factorial structure, growth processes, mean and covariance structures) applied equally to all members of

this group. In actuality, this LEP sample was very diverse with respect to ethnic, linguistic, and cultural characteristics. The two largest ethnic groups were Hispanics ( $n = 451$ ) and Asian/Pacific Islanders ( $n = 254$ ).<sup>9</sup> The remaining 194 LEP students included relatively small groups of non-Hispanic Whites, non-Hispanic Blacks, and American Indians/Alaskan Natives. The obvious question was whether the model captured in Figure 5, which fit the full sample data well, would be invariant across these ethnic groups. Sample size requirements for SEM analyses, as well as the desirability of relative group homogeneity, made it advisable to test for equivalence of parameter estimates only across the two largest LEP groups, Asians and Hispanics.

The process of evaluating multigroup invariance of measurement and structural model components typically involves testing a series of hierarchically nested and increasingly restrictive models, in which cross-group equality constraints are progressively imposed on key parameters. At each step, the fit of the more restricted model is compared against the fit of the less restricted model, and a nonsignificant chi-square difference (likelihood ratio) test is usually taken as evidence of invariance.<sup>10</sup> Because the chi-square difference test ( $\Delta_{\chi^2}$ ) is also sensitive to sample size, Cheung and Rensvold (2002) have argued that changes in goodness-of-fit indices should be used to assess differences in fit between pairs of nested models. The results of their simulation study indicated that change in CFI ( $\Delta_{CFI}$ ) is a robust index of invariance, and they proposed a  $\Delta_{CFI}$  cutoff value of .01 (i.e., a drop in CFI in the more restricted model, compared with the baseline model, equal to or smaller than .01 should lead to the conclusion that the invariance hypothesis cannot be rejected). Thus, in the present study acceptance of the multigroup invariance hypothesis required both a nonsignificant  $\Delta_{\chi^2}$  and a  $\Delta_{CFI}$  equal to or less than .01.

*Separate group analyses.* The final conditional LGM model was first estimated for each group separately. The fit of the model for the Hispanic sample was quite satisfactory,  $\chi^2(150, N = 451) = 355.490$ , CFI = .934, RMSEA = .055, SRMR = .064

<sup>8</sup> In an effort to identify other control variables that might predict L1 or L2 proficiency or the two growth factors, a model with five additional covariates was specified and estimated. These covariates included home variables (i.e., parental educational aspirations) and school-based variables either available in the NELs database (i.e., percentage of students receiving free lunch, percentage of minority students) or computed from original NELs variables (i.e., teachers' attitudes and morale, student behavior problems). The overall pattern of results for this expanded model was remarkably similar to that of the model displayed in Figure 5, but the more complex model had a poorer fit, and none of the additional variables was a significant predictor of rate of improvement in English reading; thus, those additional five covariates were not included in the final conditional model.

<sup>9</sup> Henceforth, I will refer to this group as *Asians* not only for reasons of simplicity but also because only 7 students in this group identified themselves as Pacific Islanders.

<sup>10</sup> The standard chi-square difference test cannot be used to compare the fit of nested models when the Satorra-Bentler or the Yuan-Bentler scaled test statistic is used to evaluate the fit of those models because such difference is not distributed as  $\chi^2$  (Satorra & Bentler, 2001). Step-by-step instructions for computing the correct  $\Delta_{\chi^2}$  in these cases are provided by Mplus at <http://www.statmodel.com/chidiff.html>. All chi-square difference tests reported in this article were computed following those guidelines.



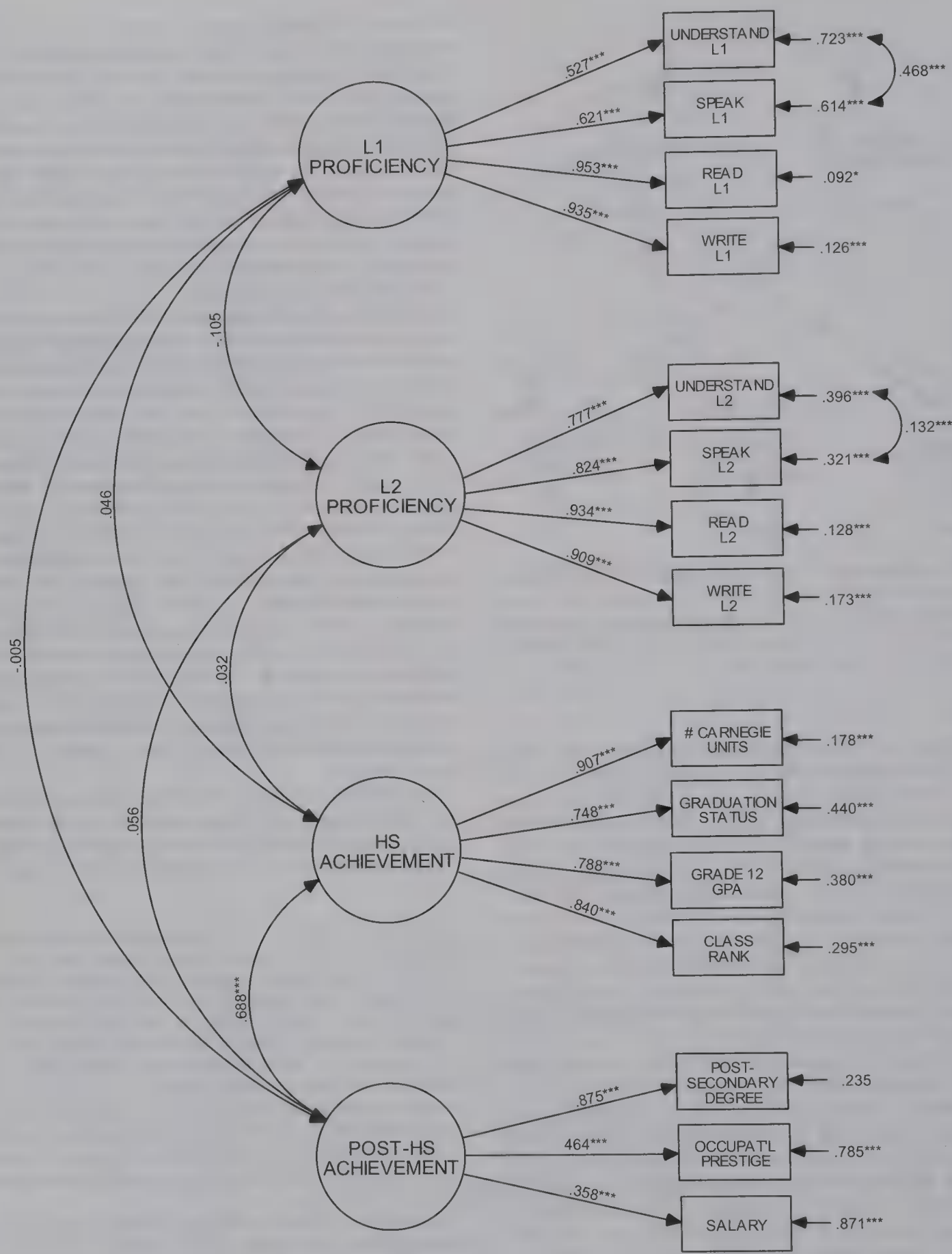
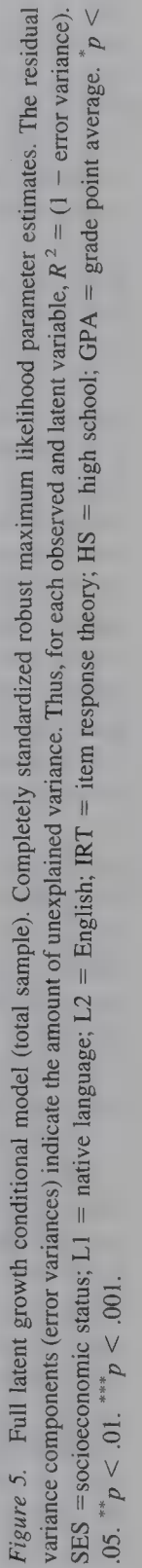


Figure 4. Confirmatory factor analysis model. Completely standardized robust maximum likelihood parameter estimates. The residual variance components (error variances) indicate the amount of unexplained variance. Thus, for each observed variable,  $R^2 = (1 - \text{error variance})$ . L1 = native language; L2 = English; HS = high school; GPA = grade point average. \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .





(Model 5, Table 2). Figure 6 shows the completely standardized estimates and their significance levels for the Hispanic students.

The general pattern of findings for this group was very consistent with the results of single group analyses. Again, both L1 and L2 proficiency significantly predicted initial L2 reading achievement scores ( $\beta = .274$ ,  $p < .001$ , and  $\beta = .185$ ,  $p < .05$ , respectively). L1 proficiency but not L2 proficiency, however, was significantly related to gains in English reading achievement over time ( $\beta = .257$ ,  $p < .005$ ), even after the effect of being schooled abroad and of SES was controlled. In fact, L1 proficiency remained the only significant predictor of English reading slope in an expanded model in which five additional family and school variables were added as covariates. In the Hispanic sample, the variables in the model explained 24.1%, 7.4%, 29.4%, and 46.3% of the variance in initial English reading level, English reading rate of growth, high school achievement, and post-high school achievement, respectively (see Figure 6).

When the same model was estimated for the Asian group, instead, there was evidence of very poor fit to the data,  $\chi^2(150, N = 254) = 641.974$ ,  $CFI = .778$ ,  $RMSEA = .114$ ,  $SRMR = .098$  (Model 8, Table 2). Particularly striking in this group was the complete absence of L1 proficiency effects on both the English reading intercept ( $\beta = -.027$ , *ns*) and the English reading slope ( $\beta = .014$ , *ns*). In addition, L2 proficiency was found to predict reading intercept ( $\beta = .244$ ,  $p < .05$ ) but not reading slope ( $\beta = .035$ , *ns*). In fact, none of the variables in the model (and none of the five additional covariates included in the expanded model) predicted English reading slope.

The separate group analyses strongly suggested that native language proficiency might play a very different role in the two groups. In the Hispanic sample, L1 proficiency was a significant predictor of both initial L2 reading level and gains in reading achievement over time. In the Asian sample, however, neither of those paths was significant. A fundamental assumption of the bilingual education approach is that L1 proficiency is critically important in the academic development of LEP students. The findings of the separate group analyses indicated that the maintenance and further development of native language skills might not serve the needs of different ethnic groups equally well. I conducted a more formal evaluation of this hypothesis by estimating separately for the two groups a model in which the key bilingual education components were missing. In Figure 6, the bilingual education assumption is modeled by the paths from L1 proficiency to the growth factors. Thus, the effect of removing those paths on model fit would provide a measure of the validity of that assumption. The two paths were deleted by fixing them to zero, and the modified model was then estimated separately for the Hispanic and Asian samples. Since this model was nested within the full LGM model displayed in Figure 5, the  $\Delta\chi^2$  and the  $\Delta CFI$  can again be used to assess the impact of model modification on fit. In the Hispanic sample, respecification of the model was associated with a significant reduction in fit,  $\Delta\chi^2(2, N = 451) = 79.052$ ,  $p < .001$ ,  $\Delta CFI = .011$ , (Model 6, Table 2). In the Asian group, on the other hand, the chi-square value of the nested model was actually lower than that of the baseline model. This produced a negative  $\Delta\chi^2$  value, which is an improper solution (Model 9, Table 2) that might be associated with either a small sample size or model misspecification (Satorra & Bentler, 2001). It is interesting that eliminating the paths from L2 proficiency to the two growth factors did not

change the fit of the model in either group,  $\Delta\chi^2(2, N = 254) = 5.218$ , *ns*,  $\Delta CFI = .000$ , for the Asian sample (Model 10, Table 2), and  $\Delta\chi^2(2, N = 451) = 6.248$ ,  $p = .044$ ,  $\Delta CFI = .002$ , for the Hispanic sample (Model 7, Table 2). Although the Hispanic group's  $\Delta\chi^2$  barely reached statistical significance, the  $\Delta CFI$  index suggests that differences between the two models are effectively negligible. The noninvariance of the full LGM model across the two ethnic groups called for a systematic evaluation aimed at identifying the nonequivalent parameters.

*Growth invariance.* The simple two-factor latent growth model presented in Figure 3 was estimated simultaneously for the two groups (Model 11, Table 2), and the fit of this unrestricted model was used as a baseline against which to assess the fit of more restricted models. The fit of this model was excellent,  $\chi^2(2, N = 705) = 2.632$ ,  $CFI = .998$ ,  $RMSEA = .030$ ,  $SRMR = .016$ . This confirmed that for both groups, the same functional form of change (i.e., a linear growth trajectory) described well the pattern of intraindividual progress in reading that occurred between Time 1 and Time 3. Increasingly more restrictive nested models were then specified and fitted to the data in order to assess the invariance of specific growth parameters. As shown in Table 2, in each case the null hypothesis of equivalence could not be rejected. When the means of the growth factors were constrained equally across the two groups (Model 12, Table 2), there was no decrement in fit relative to Model 11,  $\Delta\chi^2(2, N = 705) = 4.018$ , *ns*,  $\Delta CFI = .008$ . This implies the absence of significant group differences in initial reading levels and in rate of reading growth across time. In the next model (Model 13, Table 2), equality constraints were imposed on the variances and covariances of the growth factors while the constraints on the means were maintained. Again, there was no significant drop in fit relative to Model 12,  $\Delta\chi^2(3, N = 705) = 3.529$ , *ns*,  $\Delta CFI = .002$ . These findings support the conclusion that individual differences in initial reading scores and in the rate of reading change were essentially the same in the two groups, and there were no group differences in the relation between the two growth factors.

*Measurement invariance.* Inspection of the model estimation output suggested that a possible source of group noninvariance might be found in the measurement part of the model. Thus, the four-factor CFA model (see Figure 4) was estimated simultaneously for the two groups (Model 14, Table 2), with no equality constraints imposed on the parameters, in order to determine whether the factor structure captured in the measurement model was invariant across those groups. The poor fit of this model,  $\chi^2(164, N = 705) = 580.206$ ,  $CFI = .889$ ,  $RMSEA = .085$ ,  $SRMR = .070$ , indicated inadequate cross-cultural configural invariance. The same factor structure did not hold for Asian and Hispanic students, which means that the underlying constructs were conceptualized in different ways by the two groups. In other words, the same set of observed indicators did not reflect the same underlying constructs in the two cultural groups. Failure to demonstrate configural invariance renders all subsequent group comparisons uninterpretable and meaningless (Ployhart & Oswald, 2004; Vandenberg, 2002; Vandenberg & Lance, 2000).

In a measurement model that includes more than one factor, it is difficult to determine the source(s) of configural nonequivalence. Rensvold (personal communication, cited in Vandenberg, 2002) indicated that constructs that fail the configural invariance test could be identified by evaluating them one at a time (see Cole,

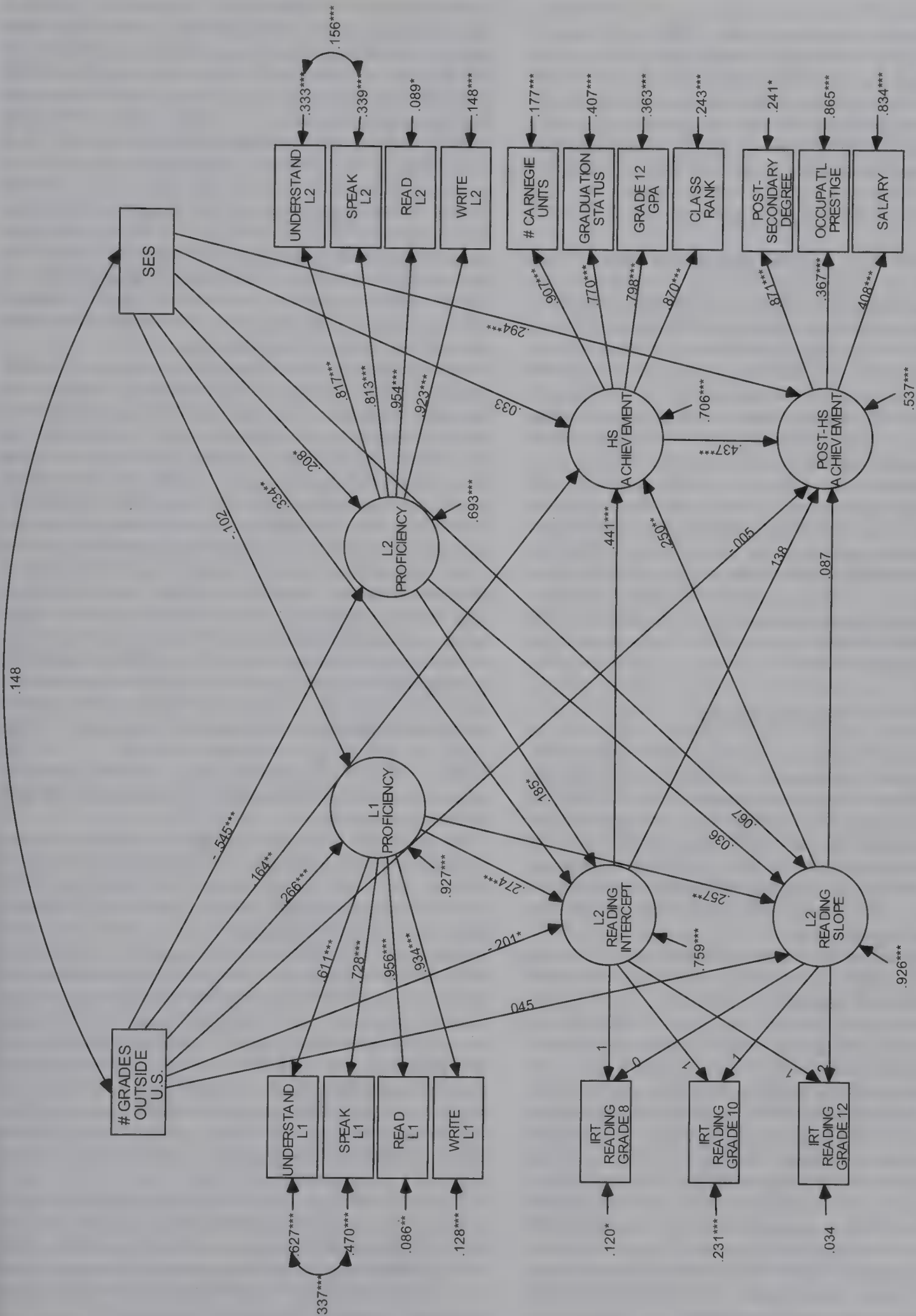


Figure 6. Full latent growth conditional model (Hispanic group). Completely standardized robust maximum likelihood parameter estimates. The residual variance components (error variances) indicate the amount of unexplained variance. Thus, for each observed and latent variable,  $R^2 = (1 - \text{error variance})$ . SES = socioeconomic status; L1 = native language; L2 = English; IRT = item response theory; HS = high school; GPA = grade point average \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .



Bedeian, & Field, 2006 for a recent example). When this strategy was applied to the measurement model in the present research, clear evidence of noninvariance was found only for the L1 proficiency construct. Rensvold, however, also pointed out that a serious limitation of this approach is that constructs may hold different meanings across groups when they are tested alone compared with when they are embedded instead in a "conceptual neighborhood" (as cited in Vandenberg, 2002, p. 147) with other constructs.

### Discussion

Latent growth curve modeling procedures were used to test some key theoretical assumptions of bilingual education. In particular, the central question examined was whether proficiency in one's native language would predict the development of English reading skills, the sine qua non for satisfactory academic achievement and positive long-term outcomes. The hypothesized relations among those constructs were incorporated into the latent growth model diagrammed in Figure 5, which included three critical paths: L1 proficiency  $\rightarrow$  English reading growth  $\rightarrow$  high school achievement  $\rightarrow$  post-high school accomplishments. The results of model estimation procedures showed a good fit to the full LEP sample and to the Hispanic subsample data. On the other hand, a substantial discrepancy was found between the observed and the model-implied covariance matrices in the Asian subsample. There was evidence that measures of key constructs did not function equivalently in the two cultural groups. Thus, following an examination of the Hispanic subsample findings, which are consistent with the basic hypotheses tested in this study, I will discuss possible reasons for group noninvariance.

### Hispanic Sample

Although over 460 languages are spoken by LEP students, Spanish is the native language for more than 79% of them (Kindler, 2002). Improving educational outcomes for the large and rapidly growing Hispanic LEP population is a national priority and remains a challenge. The results of the present study suggest that encouraging the maintenance and development of native language skills might facilitate the acquisition of English literacy, which in turn predicts positive academic and postsecondary outcomes in this language minority group.

The central finding of the Hispanic sample analyses was that self-reported L1 proficiency predicted significantly both initial values of L2 reading achievement and rate of improvement in L2 reading across 4 school years. By contrast, self-reported L2 proficiency predicted starting values but not reading trajectories. In fact, L1 proficiency turned out to be the only significant predictor of rate of improvement in English reading. This path was significant even after controlling for the contribution of the two background variables and of L2 proficiency. Despite direct negative paths from number of grades completed abroad to both self-reported L2 proficiency and English reading intercept, number of grades in non-U.S. schools positively predicted both English reading intercept and English reading slope through its association with L1 proficiency, which appeared to operate as a suppressor variable in this set of relations. The pattern of coefficients that emerged from post hoc effect decomposition analyses met the criteria for suppression effects outlined by MacKinnon, Krull, and Lockwood

(2000). Although the direct effect of being schooled abroad on English reading intercept was negative and significant, its indirect effect (through L1 proficiency) was positive and significant ( $p < .01$ ). Removal of this indirect path by fixing it to zero reduced the negative direct effect of number of grades completed abroad on English reading intercept from  $\gamma = -.201, p < .05$ , to  $\gamma = -.106, ns$ . Thus, when the variance in initial English reading scores due to L1 proficiency was controlled, the predictive power of grades completed outside the United States was sharpened. These findings indicate that being schooled outside the United States is related to lower initial levels of English reading only to the extent that it does not contribute to the development of L1 proficiency. Stated differently, the development of L1 proficiency, as a result of being schooled abroad, appears to protect students from the negative effect that having attended school outside the United States would otherwise have on their L2 reading scores.

The hypothesis that L1 proficiency predicts academic achievement through the mediation of English reading growth is consistent with bilingual education theory (e.g., Thomas & Collier, 2002) and with research on cross-language transfer. This hypothesis was also supported in the present study. In post hoc analyses of the Hispanic data, the possibility that L1 proficiency might directly predict academic performance without necessarily requiring the mediation of English literacy development was tested by adding a path from L1 proficiency to academic achievement and by fixing to zero, one at a time, the paths from the two growth factors to the achievement variable. When the path from English intercept to high school achievement was removed, L1 proficiency significantly predicted high school achievement ( $\beta = .266, p < .01$ ). Adding the English reading intercept  $\rightarrow$  high school achievement path back into the model, however, considerably attenuated the predictive strength of L1 proficiency ( $\beta = .104, ns$ ). A similar test with the English reading slope reduced the L1 proficiency  $\rightarrow$  high school achievement regression coefficient from  $\beta = .260, p < .01$ , to  $\beta = .125, ns$ . These findings suggest that L1 proficiency predicts academic performance through the mediation of English literacy development.

The self-report nature of the language proficiency measures used in this research might raise questions about the validity of its findings. The results of longitudinal CFA modeling of MTMM data, however, are comforting regarding the convergent validity, discriminant validity, and temporal stability of the L2 (and, by extension, the L1) proficiency self-ratings used in this study. Partitioning of total variance into its components indicated low proportions of method variance for the L2 proficiency measure. This makes it unlikely that students' self-ratings were substantially contaminated by the biasing effects of response style. Furthermore, if response style selectively distorted the ratings of some students (e.g., overreporting by high self-esteem responders), their estimates of L1 and L2 ability would be biased in the same direction and the correlation between L1 and L2 proficiency ratings would then be inflated. Yet, those measures were found to be uncorrelated in the present research.

In addition, the satisfactory discriminant validity of L2 proficiency self-ratings, established by the MTMM analyses, rules out the rival hypothesis that the pattern of results reported here might be explainable simply in terms of individual differences in a latent general intellectual ability factor. Other sources of evidence in the data point to the same conclusion. For example, students with



higher levels of general cognitive skills would be expected to report higher levels of proficiency in both their native language and in English; yet self-ratings of those two abilities were uncorrelated. Moreover, the path from L2 proficiency to reading growth was not significant in any of the subgroups or in the full sample, and the path from L1 proficiency to reading growth was significant in the full sample and in the Hispanic subsample. These findings are at odds with the hypothesis of a generalized intelligence effect.

### *Multigroup Noninvariance*

The results of the present study suggest that the same educational model might not be applicable to all groups of LEP students. In particular, there is evidence that at least some of the measures included in the model reflect different conceptualizations of the underlying constructs by Hispanic and Asian participants. The generally poor fit of the structural model in the Asian sample and the inability of L1 proficiency to predict growth in English reading in this group suggest that the fundamental assumptions of bilingual education theory might not hold for Asian LEP students. It is possible, however, that L1 proficiency plays an important role also in the English reading growth and general academic development of Asian LEP students, but such influence could not manifest itself because the observed variables included in the model did not adequately represent those students' conceptions of language proficiency. This possibility cannot be ruled out without first testing the same theoretical model with measuring instruments that better reflect the content domain of the constructs as they are conceptualized by Asian students. Thus, the failure to demonstrate configural cultural invariance of a measurement model holds heuristic interest in its own right insofar as it might stimulate investigations of cross-cultural differences in construct conceptualization (e.g., Cheung, Murrmann, Murrmann, & Becker, 2004; Vandenberg, 2002).

In the present study, group differences in the way the constructs were conceptualized might have been intensified by the fact that the two samples included in the multigroup analyses differed greatly in terms of within-group cultural and linguistic homogeneity. Of the 200 Asian students who responded to the question, "What was the first language you learned to speak when you were a child?" (NELS variable "bys18"), 31.5% answered "Chinese" (which includes many different spoken dialects), 14% answered "Filipino," 10.5% answered "Korean," and the rest chose "other." On the other hand, 90.7% of the Hispanic group answered "Spanish" in response to the same question. Moreover, when the Hispanic students were asked to identify the ethnic category that best described their background, 70% of them chose "Mexican." In the Asian sample, instead, only 27.7% (the largest group) answered "Chinese," followed by "Southeast Asian" (25.3%) and by eight other ethnicities. The higher the level of ethnic heterogeneity, especially when combined with marked linguistic diversity, the less likely it is that the same conceptual representation of constructs will be shared by the different subgroups.

The finding that in the Asian sample L1 proficiency, at least as conceptualized in this study, was not at all predictive of either starting levels or of rate of improvement in English reading is also consistent with the well-documented difficulty of transferring language skills from a nonalphabetic to an alphabetic language. As noted earlier, the most frequently spoken language in this group

was Chinese, a nonalphabetic language. Bialystok, McBride-Chang, and Luk (2005) found that although phonological awareness transferred from L1 to L2 in Chinese-speaking children who were learning English, such cross-language transfer did not occur for word decoding ability, a language-specific skill whose transfer is impeded by the absence of a common writing system. Similar cross-language transfer difficulties have been documented not only for young children but also for college-age English language learners who speak a nonalphabetic L1 (Akamatsu, 2003; Holm & Dodd, 1996; Wang & Koda, 2005; Wang et al., 2003).

### *Concluding Comments*

National educational surveys offer many advantages to secondary analysts: representative samples, longitudinal data, large sample sizes, carefully implemented complex sample designs, and rich data sets that include thousands of variables. On the negative side, the variables available for analysis may be inadequate to answer specific research questions. An acknowledged weakness of the present study is the unavailability of valid information about the specific type and duration of L1 and L2 support programs received by LEP students, which would have permitted a more direct test of the bilingual education model. Language support variables are available in the NELS data set and were included as covariates in an earlier version of the structural model diagrammed in Figure 5, but it became clear that their lack of validity severely limited their usefulness. For example, at base year, parents were asked, "Is your eighth grader currently enrolled in a . . . [b]ilingual or bicultural education program" (NELS variable "byp49a"). When parents responded "Yes," only 28% of their children agreed. This disconcerting lack of agreement was probably related to the frequently indiscernible nature of the L1 support programs implemented in the schools. As mentioned earlier, Rossell (2003) argued that widespread misunderstanding about bilingual education and its educational alternatives is shared by students, parents, and teachers. The NELS data set certainly provides evidence that supports this conclusion.

On the other hand, the central focus of this study was on the prediction of English reading trajectories and academic outcomes, not on the prediction of L1 proficiency. Thus, despite the inability to model L1 and L2 support measures, it can still be concluded that L1 proficiency, regardless of how it develops in Hispanic LEP students (i.e., as a result of L1 support services, of being schooled abroad, of cross-linguistic transfer effects, or of as yet unidentified factors), forecasts a favorable rate of growth in English reading skills, which is then predictive of successful short-term (i.e., high school academic achievement) and long-term (i.e., postsecondary educational and occupational attainment) outcomes.

As indicated earlier, transcript information and data for all waves were available for only 30% of the LEP students identified with the selection procedures diagrammed in Figure 1. As a result, one might wonder whether the findings of this study can be generalized to the full LEP sample. Although there were no significant differences in L1 proficiency ratings at base year between the LEP students selected for this study and the LEP students excluded because of incomplete data, there were significant differences on some dimensions. In particular, excluded LEP students exhibited significantly higher drop-out rates, lower SES, and lower self-reported English proficiency at base year than those included



in the sample. Excluded LEP students were also proportionately less likely to be Asian (though equally likely to be Hispanic) than were included LEP students. The critical question, however, is whether the latent growth model tested here, as well as its specific components (i.e., growth process, factorial structure, and structural relations among constructs), had a good fit to the initial LEP sample's data. With respect to this issue, the results of secondary analyses are encouraging. From the initial LEP sample, 1,447 students, who had been excluded because transcripts and/or data from all waves were not available, filled out the first questionnaire. The growth, measurement, and full LGM models were estimated in this subsample after the post-high school achievement factor was removed because of excessive missing data. In each case, model fit was found to be as good as or better than the fit of the same models estimated in the sample tested in this study. Once again, separate model estimation analyses for each ethnic group of excluded LEP students yielded evidence of good fit for Hispanic students and poorer fit for Asian students.

At the same time, the ability to apply these findings to the larger population of LEP students in U.S. schools is undermined by the fact that individual LEP students were declared ineligible for the NELS study if, in the opinion of the school staff, their English literacy was not sufficiently developed to enable them to read and understand the data collection instruments (Spencer, Frankel, Ingels, Rasinski, & Tourangeau, 1990). Thus the range of L2 literacy in this sample is narrower than the range found in the general LEP population. This restriction of range necessarily limits the generalizability of the findings, which can be applied only to LEP students whose English language skills are reasonably well developed. It is hoped that the policy changes in the NCES' approach to large-scale assessment instituted after the NELS data were collected and the renewed commitment to increase participation of LEP students in NCES studies (Olson & Goldstein, 1997) will make it possible for future research to determine the extent to which the model tested in this study can be applied to LEP students with lower initial levels of English proficiency.

With these caveats, the findings of the present research suggest that at least in Hispanic LEP students, native language proficiency is an important predictor of English reading trajectories. Through its association with reading growth, L1 proficiency also forecasts both short-term high school achievement and distal educational and occupational attainment. Confidence in these conclusions is strengthened by the stability of the findings. A model in which the possible confounding influence of various family and school characteristics was statistically controlled by including them as additional covariates yielded a very similar pattern of results. The extent to which existing bilingual education programs promote the development of L1 literacy and the effectiveness of these programs for particular language-minority groups need to be established by methodologically sound educational research. The findings reported here, however, point to the value of maintaining and further developing first-language proficiency for Hispanic (predominantly Mexican American) students, which is the necessary theoretical precondition for exploring the practical efficacy of specific educational interventions.

## References

- Akamatsu, N. (2003). The effects of first language orthographic features on second language reading in text. *Language Learning*, 53, 207-231.
- Asparouhov, T. (2004). *Stratification in multivariate modeling* (Mplus Web Note 9). Retrieved October 26, 2007, from [www.statmodel.com/examples/webnote.shtml](http://www.statmodel.com/examples/webnote.shtml)
- Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling*, 12, 411-434.
- August, D., & Hakuta, K. (Eds.). (1997). *Improving schooling for language-minority children*. Washington, DC: National Academy Press.
- August, D., & Hakuta, K. (Eds.). (1998). *Educating language-minority children*. Washington, DC: National Academy Press.
- Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, 107, 238-246.
- Bentler, P. M. (1995). *EQS structural equations program manual*. Encino, CA: Multivariate Software.
- Bialystok, E., Luk, G., & Kwan, E. (2005). Bilingualism, biliteracy, and learning to read: Interactions among languages and writing systems. *Scientific Studies of Reading*, 9, 43-61.
- Bialystok, E., McBride-Chang, & Luk, G. (2005). Bilingualism, language proficiency, and learning to read in two writing systems. *Journal of Educational Psychology*, 97, 580-590.
- Bilingual Education Act, P.L. No. 100-297, § 7001, 7002, 7003, 102 Stat. 130 (1988). Text from *United States Public Laws*. Available from LexisNexis Congressional at <http://www.lexisnexis.com>
- Blanche, P., & Merino, B. J. (1989). Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning*, 39, 313-340.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York: Guilford.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- Cheung, G. W., Murrmann, K. F., Murrmann, S. K., & Becker, C. (2004). Noninvariant measurement versus traditional approaches for studying cultural differences: A case of service expectations. *Journal of Hospitality & Tourism Research*, 28, 375-390.
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, 9, 233-255.
- Chou, C-P., & Bentler, P. M. (1995). Estimates and tests in structural equation modeling. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 37-55). Thousand Oaks, CA: Sage.
- Cole, M. S., Bedeian, A. G., & Field, H. S. (2006). The measurement equivalence of web-based and paper-and-pencil measures of transformational leadership: A multinational test. *Organizational Research Methods*, 9, 339-368.
- Conway, J. M., Lievens, F., Scullen, S. E., & Lance, C. E. (2004). Bias in the correlated uniqueness model for MTMM data. *Structural Equation Modeling*, 11, 535-559.
- Cummins, J. (1999). Alternative paradigms in bilingual education research: Does theory have a place? *Educational Researcher*, 28, 26-, 32, 41.
- Curtin, T. R., Ingels, S. J., Wu, S., & Heuer, R. (2002). *National Education Longitudinal Study of 1988: Base-year to fourth follow-up data file user's manual* (NCES 2002-323). Washington, DC: National Center for Education Statistics.
- Delgado, P., Guerrero, G., Goggin, J. P., & Ellis, B. B. (1999). Self-assessment of linguistic skills by bilingual Hispanics. *Hispanic Journal of Behavioral Sciences*, 21, 31-46.
- Dickinson, D. K., McCabe, A., Clark-Chiarelli, N., & Wolf, A. (2004). Cross-language transfer of phonological awareness in low-income Spanish and English bilingual preschool children. *Applied Psycholinguistics*, 25, 323-347.
- Durgunoğlu, A. Y. (2002). Cross-linguistic transfer in literacy development and implications for language learners. *Annals of Dyslexia*, 52, 189-204.



- Durgunoğlu, A. Y., Nagy, W. E., & Hancin-Bhatt, B. J. (1993). Cross-language transfer of phonological awareness. *Journal of Educational Psychology, 85*, 453–465.
- Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*, 430–457.
- Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling, 12*, 343–367.
- Geva, E., & Siegel, L. (2000). Orthographic and cognitive factors in the concurrent development of basic reading skills in two languages. *Reading and Writing, 12*, 1–30.
- Gold, M. S., Bentler, P. M., & Kim, K. H. (2003). A comparison of maximum-likelihood and asymptotically distribution-free methods of treating incomplete nonnormal data. *Structural Equation Modeling, 10*, 47–79.
- Gottardo, A., Yan, B., Siegel, L. S., & Wade-Woolley, L. (2001). Factors relating to English reading performance in children with Chinese as a first language: More evidence of cross-language transfer of phonological processing. *Journal of Educational Psychology, 93*, 530–542.
- Hakuta, K., & D'Andrea, D. (1992). Some properties of bilingual maintenance and loss in Mexican background high-school students. *Applied Linguistics, 13*, 72–99.
- Holm, A., & Dodd, B. (1996). The effect of first written language on the acquisition of English literacy. *Cognition, 59*, 119–147.
- Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.
- Hu, L. T., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424–453.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.
- Ingels, S. J., Scott, L. A., Rock, D. A., Pollack, J. M., & Rasinski, K. (1994). *NELS:88 first: Follow-up final technical report* (NCES 94–632). Washington, DC: National Center for Education Statistics.
- Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling, 8*, 325–352.
- Kaufman, P., & Rasinski, K. A. (1991). *Quality of the responses of eighth-grade students in NELS:88* (NCES 91–487). Washington, DC: National Center for Education Statistics.
- Kenny, D. A., & Kashy, D. A. (1992). Analysis of the multitrait-multimethod matrix by confirmatory factor analysis. *Psychological Bulletin, 112*, 165–172.
- Kindler, A. L. (2002). *Survey of the states' limited English proficient students and available educational programs and services: 2000–2001 summary report*. Washington, DC: U.S. Department of Education, Office of English Language Acquisition, Language Enhancement and Academic Achievement for Limited English Proficient Students.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.
- Lance, C. E., Noble, C. L., & Scullen, S. E. (2002). A critique of the correlated trait–correlated method and correlated uniqueness models for multitrait–multimethod data. *Psychological Methods, 7*, 228–244.
- Lau v. Nichols, 414 U.S. 563 (1974).
- MacKinnon, D. P., Krull, J. L., & Lockwood, C. M. (2000). Equivalence of the mediation, confounding, and suppression effect. *Prevention Science, 1*, 173–181.
- Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*, 320–341.
- McLaughlin, D. H., & Cohen, J. (1997). *NELS:88 survey item evaluation report* (NCES 97–052). Washington, DC: National Center for Education Statistics.
- Muthén, B., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 267–316). Washington, DC: American Sociological Association.
- Muthén, L., & Muthén, B. (2006). *Mplus user's guide* (4th ed.). Los Angeles, CA: Muthén & Muthén.
- Nakao, K., & Treas, J. (1994). Updating occupational prestige and socioeconomic scores: How the new measures measure up. *Sociological Methodology, 24*, 1–72.
- Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods, 6*, 328–362.
- Olson, J. F., & Goldstein, A. A. (1997). *The inclusion of students with disabilities and limited English proficient students in large-scale assessments: A summary of recent progress* (NCES 97–482). Washington, DC: National Center for Education Statistics.
- Ployhart, R. E., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods, 7*, 27–65.
- Portes, A., & Hao, L. (1998). *E pluribus unum: Bilingualism and loss of language in the second generation*. *Sociology of Education, 71*, 269–294.
- Rock, D. A., & Pollack, J. M. (1991). *Psychometric report for the NELS:88 base year test battery* (NCES 91–468). Washington, DC: National Center for Education Statistics.
- Rock, D. A., & Pollack, J. M. (1995). *Psychometric report for the NELS:88 base year through second follow-up* (NCES 95–382). Washington, DC: National Center for Education Statistics.
- Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15*, 1–20.
- Rossell, C. H. (2003). *Policy matters in teaching English language learners: New York and California* (Report No. UDS-117). New York: ERIC Clearinghouse on Urban Education Institute for Urban and Minority Education.
- Rossell, C. H., & Baker, K. (1996). The educational effectiveness of bilingual education. *Research in the Teaching of English, 30*, 7–74.
- Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika, 66*, 507–514.
- Savalei, V., & Bentler, P. M. (2005). A statistically justified pairwise ML method for incomplete nonnormal data: A comparison with direct ML and pairwise ADF. *Structural Equation Modeling, 12*, 183–214.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177.
- Shameem, N. (1998). Validating self-reported language proficiency by testing performance in an immigrant community: The Wellington Indo-Fijians. *Language Testing, 15*, 86–108.
- Slavin, R. E., & Cheung, A. (2003). *Effective reading programs for English language learners: A best-evidence synthesis* (Report No. 66). Baltimore, MD: Johns Hopkins University.
- Spencer, B. D., Frankel, M. R., Ingels, S. J., Rasinski, K. A., & Tourangeau, R. (1990). *National Education Longitudinal Study of 1988: Base year sample design report* (NCES 90–463). Washington, DC: National Center for Education Statistics.
- Stapleton, L. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling, 9*, 475–502.
- Stapleton, L. M. (2006). An assessment of practical solutions for structural equation modeling with complex sample data. *Structural Equation Modeling, 13*, 28–58.
- Steiger, J. H., & Lind, J. M. (1980, May). Statistically based tests for the number of common factors. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.



- Strang, W., Winglee, M., & Stunkard, J. (1993). *Characteristics of secondary-school-age language minority and limited English proficient youth: Final analytic report*. Retrieved October 29, 2007, from the National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs Web site, <http://www.ncela.gwu.edu/pubs/siac/secondary/>
- Thomas, W., & Collier, V. P. (2002). *A national study of school effectiveness for language minority students' long-term academic achievement*. Santa Cruz, CA, and Washington, DC: Center for Research on Education, Diversity, and Excellence.
- Tomarken, A. J., & Waller, N. G. (2003). Potential problems with well fitting models. *Journal of Abnormal Psychology, 112*, 578–598.
- Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*, 139–158.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4–70.
- Verhoeven, L. (1994). Transfer in bilingual education development: The linguistic interdependency hypothesis revisited. *Language Learning, 44*, 381–415.
- Wang, M., & Koda, K. (2005). Commonalities and differences in word identification skills among learners of English as a second language. *Language Learning, 55*, 71–98.
- Wang, M., Koda, K., & Perfetti, C. A. (2003). Alphabetic and nonalphabetic L1 effects in English word identification: A comparison of Korean and Chinese English L2 learners. *Cognition, 87*, 129–149.
- Yeung, A. S., Marsh, H. W., & Suliman, R. (2000). Can two tongues live in harmony: Analysis of the National Education Longitudinal Study of 1988 (NELS88) longitudinal data on the maintenance of home language. *American Educational Research Journal, 37*, 1001–1026.
- Yuan, K. H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 30*, 165–200.

Received February 16, 2006

Revision received August 30, 2007

Accepted September 27, 2007 ■

# Growth in Working Memory and Mathematical Problem Solving in Children at Risk and Not at Risk for Serious Math Difficulties

H. Lee Swanson  
University of California, Riverside

Olga Jerman  
Frostig School

Xinhua Zheng  
University of California, Riverside

The influence of cognitive growth in working memory (WM) on mathematical problem solution accuracy was examined in elementary school children ( $N = 353$ ) at risk and not at risk for serious math problem solving difficulties. A battery of tests was administered that assessed problem solving, achievement, and cognitive processing (WM, inhibition, naming speed, phonological coding) in children in 1st, 2nd, and 3rd grade across 3 testing waves. The results were that (a) children identified as at risk for serious math problem solving difficulties in Wave 1 showed less growth rate and lower levels of performance on cognitive measures than did children not at risk; (b) fluid intelligence and 2 components of WM (central executive, visual-spatial sketchpad) in Wave 1 (Year 1) predicted Wave 3 word problem solving solution accuracy; and (c) growth in the central executive and phonological storage component of WM was related to growth in solution accuracy. The results support the notion that growth in WM is an important predictor of children's problem solving beyond the contribution of reading, calculation skills, and individual differences in phonological processing, inhibition, and processing speed.

**Keywords:** working memory, math disabilities, problem solving, phonological processing, executive processing

Word problems are an important part of mathematics programs in elementary schools. This is because word problems help students apply formal mathematical knowledge and skills to real world situations. Much of the evidence indicates that word problem performance improves as children gain greater ability in (a) understanding underlying arithmetic operations (e.g., Rasmussen & Bisanz, 2005), (b) distinguishing between types of word problems on a basis of mathematical operations (e.g., Fayol, Abdi, & Gombert, 1987; Rittle-Johnson, Siegler, & Alibali, 2001), and (c) an effective use of strategies (e.g., Geary, Hoard, Byrd-Craven, & Desoto, 2004; Siegler, 1988). Developmental changes in skills in mathematics do not provide a complete account, however, of

age-related changes in word problem solving. There is evidence that suggests the need for more general cognitive processes, that is, processes nonspecific to mathematics. For example, solving a word problem, such as "15 dolls are for sale. 7 dolls have hats. The dolls are large. How many dolls do not have hats?" involves the development of a variety of mental activities (Barrouillet & Lépine, 2005). Children must access prestored information (e.g., 15 dolls), access the appropriate algorithm (15 minus 7), and apply problem solving process to control its execution (e.g., ignore the irrelevant information). Given the multistep nature of word problems, it seems plausible that working memory (WM) plays a major role in solution accuracy. WM is defined as a processing resource of limited capacity, involved in the preservation of information while simultaneously processing the same or other information (Baddeley & Logie, 1999; Engle, Tuholski, Laughlin, & Conway 1999; Miyake, 2001).

Although there are several models of WM, Baddeley's multi-component model has often been used to explore the role of WM on mathematical problem solving (e.g., Swanson & Beebe-Frankenberg, 2004). Baddeley and Logie (1999) described WM as a limited capacity central executive system that interacts with a set of two passive store systems used for temporary storage of different classes of information: the speech-based phonological loop and the visual-spatial sketchpad. The phonological loop is responsible for the temporary storage of verbal information; items are held within a phonological store of limited duration, and the items are maintained within the store via the process of articulation (inner vocalization). The visual-spatial sketchpad is responsible for the storage of visual-spatial information over brief periods and

---

H. Lee Swanson, Graduate School of Education, University of California, Riverside; Olga Jerman, Frostig School, Pasadena, California; Xinhua Zheng, Department of Educational Psychology, University of California, Riverside.

This article is based on a 3-year longitudinal study funded by the U.S. Department of Education, Cognition and Student Learning (USDE R305H020055), Institute of Education Sciences awarded to H. Lee Swanson. We thank Georgia Doukas, Diana Dowds, Rebecca Gregg, Krista Healy, Crystal Howard, James Lyons, Kelly Rosston, and Leilani Sáez for data collection and/or task development; the Colton School District, Tri City Christian Schools, and the Frostig School; and Margaret Beebe-Frankenberg and Bev Hedin who directed and managed data collection and school schedules. This article does not necessarily reflect the views of the U.S. Department of Education or the school districts.

Correspondence concerning this article should be addressed to H. Lee Swanson, Graduate School of Education, University of California, Riverside, CA 92521. E-mail: lee.swanson@ucr.edu



plays a key role in the generation and manipulation of mental images. Both storage systems are in direct contact with the central executive system. The central executive system is considered to be primarily responsible for coordinating activity within the cognitive system but also devotes some of its resources to increase the amount of information that can be held in the two subsystems (Baddeley & Logie, 1999). A recent formulation of the model (Baddeley, 2000) also includes a temporary multimodal storage component called the *episodic buffer*.

Although WM is a fundamental component in many current theories of children's problem solving (in addition to learning, comprehension, and reasoning), no longitudinal studies (to our knowledge) have explicitly isolated those components of WM most directly related to growth in problem solving, especially in children at risk for word problem solving difficulties. Because WM is made up of three components (central executive, phonological loop, and visual-spatial sketchpad), the question arises as to whether WM as a whole operates on problem solving or whether a certain component is more important. Thus, although there is a strong interrelationship among these components, we sought in the present study to determine whether growth in certain components of WM is related to growth in word problem solving. In this study, we attempted to break out the contributions of WM by focusing on short-term memory (STM) measures, verbal WM measures, and visual-spatial WM measures. In elaborating on the distinction between STM and WM, Cowan (1995) emphasized the role of attentional processes. WM is depicted as a subset of items of information stored in STM that are in turn submitted to limited attentional control processing (see also Engle et al., 1999). This assumes that when the contents of STM are separated from WM what is left of WM is controlled attention or processing related to the central executive system (also referred to in this article as *executive processing* or *the central executive component of WM*). Consequently, to understand the impact of WM to problem solving in terms of executive processing the influence of STM must be partialled out.

### Models of WM and Word Problem Solving

How might growth in WM mediate age-related and individual differences in word problem solving? We consider three models as an explanation of the role of WM in individual and age-related problem solving performance in children: one focuses on the child's knowledge base for arithmetical calculations and components of word problems; another focuses on the storage components of WM, primarily the phonological loop; and the third focuses on the central executive system. The models are not necessarily exclusive of one another (each process can contribute important variance to problem solving to some degree) but suggest that some processes are more important than others.

#### *Knowledge Base*

The first model considers whether age-related differences in the child's knowledge base play a major role in mediating the influence of WM on problem solving. Several capacity models suggest that WM represents an activated portion of declarative long-term memory (LTM) (J. R. Anderson, Reder, & Lebiere, 1996; Cantor & Engle, 1993). That is, WM capacity influences the amount of

resources available to activate knowledge (see Conway & Engle, 1994, for a review of this model). Baddeley and Logie (1999) stated that a major role of WM "is retrieval of stored long-term knowledge relevant to the tasks at hand, the manipulation and recombination of material allowing the interpretation of novel stimuli, and the discovery of novel information or the solution to problems" (p. 31). They further stated that "any increase in total storage capacity beyond that of a given slave system is achieved by accessing either long-term memory (LTM) or other subsystems" (p. 37). Thus, the influence of WM performance on problem solving is related to one's ability to accurately access information (e.g., appropriate algorithm) from LTM to solve the problem. More specifically, a word problem introduces information into WM. The contents of WM are then compared with possible action sequences (e.g., associative links) in LTM (Ericsson & Kintsch, 1995). When a match is found (recognized), the contents of WM are updated and used to generate a solution. This assumption is consistent with current models of problem solving that are based on "recognize-act" models of a cognitive processor (J. R. Anderson et al., 1996; Ericsson & Kintsch, 1995).

In the present study, we assessed whether the retrievability of contents in LTM such as math skills, as well as domain-specific knowledge of the propositions found in word problems, mediates WM and problem solving. These LTM propositions are related to accessing numerical, relational, question, and extraneous information, as well as accessing the appropriate operations and solution algorithms (Hegarty, Mayer, & Monk, 1995; Mayer & Hegarty, 1996; Swanson, Cooney, & Brock, 1993). In the present study, we assumed that the contribution of LTM could be tested by partialing its influence from the correlations between WM and problem solving. Clearly, we only sampled some aspects of LTM, but we assumed that the influence of WM on problem solution accuracy should be eliminated when measures related to LTM (e.g., knowledge of the propositions found in word problems) are partialled from the analysis.

#### *Phonological System*

The second model assumes that the individual and age-related influence of WM on children's mathematic problem solving is primarily moderated by STM storage, the phonological system in particular. Because mathematical word problems are presented in a text format and the decoding and comprehension of text draws on the phonological system (see Baddeley, Gathercole, & Papagno, 1998, for a review), individual differences and age-related differences on problem solving tasks can be attributed to the phonological loop. Several studies assume that STM measures capture a subset of WM performance, the utilization and/or operation of the phonological loop (for a comprehensive review, see Dempster, 1985; Gathercole, 1998). This is because successful performance on STM measures draws on two major components of the phonological loop: a speech-based phonological input store and a rehearsal process (see Baddeley, 1986, for review). Research to date on STM indicates that young children rehearse less and perform more poorly on tasks requiring the short-term retention of order information when compared with older children (Ornstein, Naus, & Liberty, 1975), and children with math disabilities rehearse less when compared with children without math disabilities (e.g., see Geary, 2003, for a review). This suggests inefficient utilization of



the phonological rehearsal process (Henry & Millar, 1993). Because younger children and children with math disabilities have smaller digit spans than older children and children without math disabilities, it is possible they have basic inefficiencies in the storage of phonological input that influence higher level processing, such as comprehending and solving word problems. Developmental and individual differences in the phonological loop, therefore, might be expected to influence some aspects of problem solving, such as computing solutions to problems (Furst & Hitch, 2000). The phonological loop may be able to retain information of verbal form during ongoing calculations. Although not in the domain of math, *per se*, some studies suggest that simple short-term storage has a significant role in accounting for the relationship between WM and several cognitive abilities (Ackerman, Beier, & Boyle, 2005; Colom, Abad, Rebollo, & Shih, 2005; Colom, Flores-Mendoza, Quiroga, & Privado, 2005).

Thus, a simple version of this hypothesis states that individuals at risk for math disabilities and younger children are slower and/or less accurate at processing verbal information (numbers, letters) than average achieving children or older children, and this reduced processing on the participants' part underlies their poor WM and problem solving performance. These assumptions are consistent with a number of bottom-up models of higher order processing, such as comprehension, which view the primary task of executive processing as one of relaying the results of lower level linguistic analyses upward through the language system (Shankweiler & Crain, 1986). Phonologically analyzed information is transferred to WM storage, which in turn is then transferred (thus freeing storage for the next chunk of phonological information) upward through the processing system to promote online extraction of meaning. One of the possible reasons WM span increases as a function of age is because older children and children without math disabilities can name items more rapidly at recall than younger children or children with math disabilities. That is, increases in naming from early to the late childhood years are assumed to enhance the effectiveness of subvocal rehearsal processes and hence reduce the decay of memory items in the phonological store prior to output (Henry & Millar, 1993).

There are clear expectations in the aforementioned model: Individual and age-related changes in children's problem solving are related to the phonological loop. Mathematical proficiency follows automatically from improvements in phonological processing (i.e., because of improved storage and speed of processing information). Therefore, correlations between WM and problem solving should be significantly weakened if measures reflective of the phonological loop are partialled from the analysis. In other words, if age-related differences in problem solving performance and WM are moderated by the phonological system, then the relationship between problem solving and WM should be diminished when measures of the phonological system (e.g., phonological knowledge, speed, STM) are partialled from the analysis.

### *Central Executive System*

The third model incorporates some of the assumptions of the first two models by viewing executive processing as: (a) providing resources to lower order (phonological system) skills and (b) accessing information from LTM. However, the model also views executive processes that are independent of those skills as playing

a major role in individual differences in mathematical problem solving. That is, although individual and age-related differences in problem solving accuracy are possibly related to the retrievability of contents in LTM (e.g., knowledge of specific mathematical relations, general problem solving strategies) and the phonological loop, other activities of the executive system may also underlie the influence of WM on solution accuracy (e.g., Swanson & Ashbaker, 2000; Swanson & Sachse-Lee, 2001). For example, several other cognitive activities (e.g., see Miyake, Friedman, Emerson, Witzki, & Howerter, 2000, for a review), such as controlling subsidiary memory systems, control of encoding and retrieval strategies, attention switching during manipulation of material held in the verbal and visual-spatial systems, and suppressing irrelevant information, have been assigned to the central executive system (Baddeley, 1996; Miyake et al., 2000; Oberauer, Suss, Wilhelm, & Wittman, 2003). Thus, the third model suggests that the central executive system contributes significant variance to individual and age-related differences in problem solving. Thus, in contrast to the aforementioned models that suggest increases in knowledge base and phonological processes play the dominant role in the mediating effects of WM on children's word problem solving, the present model assumes that executive processes also play a major role in mediating problem solving performance. There are clear expectations for this model. The influence of WM on measures of problem solving follows automatically with improvements in performance related to the executive component of WM.

### *Previous Individual and Cross-Sectional Studies*

Some preliminary work has been done in addressing these three models. Studies by Swanson and colleagues (Swanson, 2004; Swanson & Beebe-Frankenberger, 2004; Swanson & Sachse-Lee, 2001) have shown that children with mathematical problem solving difficulties across different age groups perform poorly relative to controls on measures of WM. For example, Swanson and Sachse-Lee (2001) found with children ages 11 to 15 with and without math disabilities that phonological processing, verbal WM, and visual WM contributed unique variance to solution accuracy in word problems. This finding is in line with the current literature suggesting that the development of the phonological system plays an important part in accounting for individual differences in text processing. However, their results also showed that verbal WM processes played just as important a role as phonological processes (i.e., verbal WM reduced the contribution of the ability group contrast variable by approximately 63%) in accounting for the ability group differences in solution accuracy. Thus, there was weak support for the assumption that only the development of bottom-up processes (i.e., the phonological system) mediated deficits in WM processing and its influence on solution accuracy. In a follow-up analysis of Swanson and Beebe-Frankenberger's (2004) study, Swanson (2006) found that when residual variance related to STM was partialled out in the analysis, the component that best predicted problem solving was the central executive system of WM.

Unfortunately, none of the aforementioned studies have determined whether developmental increases in WM are related to developmental increases in problem solving. There have been some studies that have investigated cross-sectional differences in WM and word problem solving as a function of age (e.g., Swan-



son, 1999), however, no studies that we are aware of have investigated longitudinally whether changes in children's WM are related to changes in problem solving. Thus, despite studies showing that WM and problem solving improve as a function of age in cross-sectional studies (Swanson, 1999, 2003), the influence of growth in WM on problem solving in children has not been established. Identifying such a growth effect would be extremely valuable because it would assist in providing a parsimonious account of the ability group differences in WM found across a number of problem solving measures.

This study addressed the questions as to whether growth in problem solving is related to growth in WM and whether these changes vary as a function of children with and without serious mathematical problem solving disabilities in three elementary age groups and two ability groups (children at risk and not at risk for serious math difficulties). On the basis of the aforementioned models, we considered three possibilities: (a) Growth in the relationship between WM and problem solving is primarily mediated by task-specific knowledge and skills in mathematical calculation; (b) growth in WM and problem solving is primarily mediated by the phonological loop; or (c) growth in the executive component of WM, independent of the phonological system and resources activated in LTM, contributes unique variance to problem solving.

For this study, we assessed LTM using performance on measures of arithmetic calculation and recognition of problem solving components. We assessed the phonological system using performance on measures of naming speed, phonological awareness, and STM. We assessed executive processing using performance on WM tasks modeled after the format of Daneman and Carpenter's (1980) measure. The WM tasks were assumed to capture at least two factors of executive processing: susceptibility to interference and manipulation of capacity in the coordination of both processing and storage (e.g., Oberauer, 2002; Whitney, Arnett, Driver, & Budd, 2001). The approach used in this study (as well as in others, e.g., Engle, Cantor, & Carullo, 1992), to assess whether a particular system plays the major role in mediating differences in performance, was to remove statistically that system's influence from the analysis. In this study, the influence of the phonological system (e.g., naming speed, STM) or LTM was partialled via a hierarchical regression analysis between problem solving and WM. We reasoned that if WM and problem solving are primarily mediated by a phonological system and/or LTM, then the predictions of problem solving performance by performance on WM measures should be nonsignificant when measures related to the phonological system and LTM are entered (partialled) in the analysis. However, if growth in the central executive system (executive processing) mediates the relationship between WM and problem solving, then the correlations between these two variables will remain significant when measures of phonological processing and LTM are partialled from the analysis.

In summary, three questions directed our study.

1. Do children identified at risk or not at risk for serious math problem solving difficulties in Wave 1 vary on measures of growth in problem solving and WM across the three testing waves?
2. Is growth in WM related to growth in word problem solving?
3. Which components of WM are related to predictions of problem solving accuracy and are those predictions mediated by individual differences in knowledge base, phonological processing, and/or a central executive system?

## Method

### Participants

For the first wave of data collection, 353 children from Grades 1, 2, and 3 from a Southern California public school district and private school district participated in this study. Final selection of participants was determined by parent approval for participation and achievement scores. Of the 353 children selected, 167 were boys and 186 were girls. Gender representation was not significantly different among the three age groups,  $\chi^2(2, N = 353) = 1.15, p > .05$ . Ethnic representation of the sample was 163 Anglo, 147 Hispanic, 25 African American, 14 Asian, and 4 other (e.g., Native American and Vietnamese). The mean socioeconomic status of the sample was primarily middle class on the basis of parent education or occupation. However, the sample varied from low-middle class to upper middle class. Means and standard deviations for the variables used in this study for all testing waves are shown in Appendix A, B, and C.

The second wave of the testing 1 year later included 320 children, and the third wave of testing 2 years later included 302 children. The attrition of children who dropped out of the study was due to moving out of the school district. A comparison was made among achievement and cognitive scores, socioeconomic status, and parent income or occupation between children retained and those not retained in the study. In the Wave 1 sample, 134 children (68 boys, 66 girls) were classified at risk for serious math problem solving difficulties (SMD), and 219 children (99 boys and 120 girls) were not at risk. No significant differences emerged between the two ability groups in terms of ethnicity,  $\chi^2(5, N = 353) = 8.67, p > .05$ , or gender,  $\chi^2(1, N = 353) = 0.06, p > .05$ . With regard to attrition, however, the at-risk sample was reduced from 134 to 116 (64 boys and 52 girls) children, and the not-at-risk sample was reduced from 219 to 186 (94 boys and 92 girls) children in Wave 3. The comparison of the two groups at Wave 3 indicated that no significant differences emerged between the risk groups in terms of gender,  $\chi^2(1, N = 302) = 0.61, p > .05$ . Differences did emerge, however, between the risk groups in terms of ethnicity,  $\chi^2(5, N = 302) = 12.03, p < .05$ . Ethnic representation of the SMD ( $n = 116$ ) sample in Wave 3 was 44 Anglo, 58 Hispanic, 7 African American, 4 Asian, and 3 other (e.g., Native American and Vietnamese). In contrast, ethnic representation for the children not at risk ( $n = 186$ ) was 96 Anglo, 63 Hispanic, 14 African American, 12 Asian, and 1 other (Vietnamese). Thus, the at-risk group in Wave 3 had lower Anglo representation (only 38% of the sample) than the children not at risk (52% of the sample).

No bias was detected in child attrition except that proportionally more Hispanic children dropped out of the study. However, we did not find that gender interacted significantly with the performance on dependent measures (all  $ps > .05$ ). Therefore, gender was not considered further in the analysis. The sample size and scores as a



function of age, ability group, and testing wave for each task are shown in Appendix A, B, and C.

### *Definition of Children at Risk for SMD*

In this study, children at risk for SMD were defined as having normal intelligence on the basis of a measure of fluid intelligence (in this case, the Raven Colored Progressive Matrices test score  $\geq 85$ ; Raven, 1976) but with a mean performance below the 25th percentile (standard score of 90 or scaled score of 8) on norm-referenced measures related to (a) solving orally presented word problems and (b) digit naming fluency. The 25th percentile cutoff score on standardized achievement measures has been commonly used to identify children at risk (e.g., Fletcher et al., 1989) and, therefore, was used in this study. Classification of children at risk for SMD and not at risk for serious math difficulties was based on norm-referenced measures of computation on the Arithmetic subtest of the Wechsler Intelligence Scale for Children—Third Edition (WISC-III; Psychological Corporation, 1991) and Digit Naming Speed from the Comprehensive Test of Phonological Processing (CTOPP, Wagner, Torgesen, & Rashotte, 2000) described in the following section. Children who yielded an average scaled score less than or equal to 8 (average of both measures) were considered at risk for SMD. A scaled score of 8 was equivalent to a standard score of 90 or a percentile score of 25.

Our rationale for the above classification procedures was as follows. There are no general agreed upon criteria for defining children at risk for serious math difficulties in problem solving, especially in Grade 1 when the instruction is only beginning to address mathematical operations and learning to read. In addition, few studies have focused on math word problem solving difficulties in younger children, let alone identifying possible definitional parameters of risk. Because first graders were used in our sample in Wave 1, it was necessary in our classification of children at risk to control for the demands placed on reading and writing. In addition, because our focus was on examining problem solving skill and not arithmetic calculation, per se, for classification purposes it was necessary to rely on different measures than studies that define math disabilities by computation skill (i.e., mental computation of word problems versus paper and pencil computation of arithmetic problems). In contrast to the literature on math disabilities or reading disabilities, we assumed that children with SMD may not have skill difficulties related to arithmetic calculation or reading, per se, but may nevertheless have serious difficulties in coordinating arithmetic and language processes to solve a problem. For example, children we identified at risk had reading and math norm-referenced scores in the average range (see Appendix A, B, and C). Further, because we were interested in the reasoning processes related to problem solving, we selected a reliable measure that presented questions orally and placed problems in a verbal context (e.g., "If I have an candy bar and divide it in half, how many pieces do I have?") rather than a written computational context ("1 + 1 = ?"). Thus, we utilized the oral presentation of story problems as a criterion measure of SMD. Therefore, children who scored at or below a scaled score of 8 were selected as SMD. No doubt, this measure has been found to be a "complex mix of abilities, including quantitative reasoning . . . (a narrow ability subsumed by fluid reasoning), working memory, verbal comprehension and knowledge . . . the first choice for

interpretation of Arithmetic should likely be as a measure of fluid reasoning" (Keith, Fine, Taub, Reynolds, & Kranzler, 2006, p. 118). This interpretation was also supported in the WISC-III version (Keith & Wittam, 1997).

In selecting children with SMD at Wave 1, we also focused on a child's verbal fluency with numbers during Wave 1 testing. We assumed that number naming speed may underlie children's ability to automatically access arithmetic facts and procedures (Bull, Johnston, & Roy, 1999; Geary, Brown, & Samaranayake, 1991). We assumed that children who have quicker access to numbers (i.e., faster number fluency) would be less at risk for mental computation difficulties than children less fluent in number naming. This seemed reasonable to us based on Hitch and McAuley's (1991) finding that children with math disabilities evidenced deficits in the speed of implicit counting. Further, speed of number naming has a parallel in the reading literature where both letter naming speed and phonological knowledge are assumed to underlie reading disabilities.

As will be shown in the Results section, the SMD classification of children in Wave 1 also differentiated children's performance in Waves 2 and 3. However, it is important to note that a small number of children ( $n = 15$ ; 13.16 % of the sample) did not stay at or below the 25th percentile (scale score of 8) across testing waves. Clearly, the problem solving and digit naming speed measure were not perfectly correlated and therefore the 25th percentile cutoff does not hold for both measures. For example, standard scores for the digit naming task were not stable over the three testing waves because children's scale score moved from an 8 in Wave 1 to a 9 in Wave 3 (42% of the sample). We reran the analysis comparing at-risk and not-at-risk groups using only the WISC-III Math subtest as an indicator of ability (cutoff scores at or less than a scale score of 8). However, because the pattern in performance across measures was comparable, we used both measures to classify children at risk for SMD during Wave 1 testing.

### *Tasks and Materials*

The battery of group and individually administered tasks is described below. Experimental tasks are described in more detail than published and standardized tasks. Tasks were divided into classification, criterion, and predictor variables. Cronbach's alpha reliability coefficients for the sample were calculated for all scores across all testing waves.

### *Classification Measures*

*Fluid intelligence.* Nonverbal intelligence was assessed by the Raven Colored Progressive Matrices test (Raven, 1976). We assumed that this measure tapped components of fluid intelligence in children (see Klauer, Willmes, & Phye, 2002; Stoner, 1982). Children were given a booklet with patterns displayed on each page, each pattern revealing a missing piece. For each pattern, six possible replacement pattern pieces were displayed. Children were required to circle the replacement piece that best completed the patterns. After the introduction of the first matrix, children completed their booklets at their own pace. Patterns progressively increased in difficulty. The dependent measure (range = 0–36) was the number of problems solved correctly, which yielded a standardized score ( $M = 100$ ,  $SD = 15$ ). The sample Cronbach's



alpha coefficients on scores for Waves 1, 2, and 3 on this task were .87, .89, and .90, respectively.

*Mental computation of word problems.* This task was taken from the Arithmetic subtest of the WISC-III (Psychological Corporation, 1991). Each word problem was orally presented and was solved without paper or pencil. The dependent measure was the number of problems correct, which yielded a scaled score ( $M = 10$ ,  $SD = 2$ ). The sample Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 on this task were .74, .65, and .69, respectively.

*Digit naming speed.* The administration procedures followed those specified in the manual of the CTOPP (Wagner et al., 2000). For this task, the examiner presented participants with an array of 36 digits. Participants were required to name the digits as quickly as possible for each of two stimulus arrays containing 36 items, for a total of 72 items. The task administrator used a stopwatch to time participants on speed of naming. The sample Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 on this task were .92, .94, and .95, respectively.

### Criterion Variables

*Word problems—semantic structure varied.* The purpose of this experimental measure was to assess mental problem solving as a function of variations in the semantic structure of a word problem (see Swanson & Beebe-Frankenberger, 2004, for details on this measure). Children were orally presented a problem and were asked to calculate the answer in their head. The word problems were derived from the work of Riley, Greeno, and Heller (1983), Kintsch and Greeno (1985), and Fayol et al. (1987). There are four sets of questions. Eight questions within each set were ordered by the difficulty of responses. The dependent measure was the number of problems solved correctly. The total possible number of questions correct was 32. The sample Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 for total scores on this task were .69, .54, and .65, respectively.

### Word Problem Solving Components

*Mathematical word problem solving processes.* This experimental test assessed a child's ability to retrieve the propositions or components related to word problems (see Swanson & Beebe-Frankenberger, 2004, for details on this measure). The components assessed were derived from the earlier work of Mayer (see Mayer & Hegarty, 1996, for a review). Four booklets were administered that each contained three word problems and a series of multiple-choice questions. Problems were four sentences in length and contained two-assignment propositions (one relation, one question) and an extraneous proposition related to the solution. To control for reading problems, an examiner orally read each problem and all multiple-choice response options as the students followed along. Questions assessed the students' ability to correctly (a) identify the question proposition of each story problem, (b) identify the numbers in the propositions of each story problem, (c) identify the goals in the assignment propositions of each story problem, (d) correctly identify the operation, and (e) correctly identify the algorithm for each story problem. At the end of each booklet students were read a series of true-false questions. All statements were related to the extraneous propositions for each

story problem within the booklet. The total score possible for propositions related to question, number, goal, operations, algorithms, and true-false questions was 12. The sample Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 for total scores on this task were .97, .88, and .91, respectively.

### Arithmetic Calculation

*Arithmetic computation.* The Arithmetic subtest from the Wide Range Achievement Test—Third Edition (WRAT-III; Wilkinson, 1993) and the Wechsler Individual Achievement Test (WIAT; Psychological Corporation, 1992) were administered. The sample Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 for the WRAT-III were .78, .84, and .71, respectively. For the WIAT, the sample Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 were .89, .86, and .76, respectively.

*Computation fluency.* This test was adapted from the Test of Computational Fluency (Fuchs, Fuchs, Eaton, Hamlett, & Karns, 2000). The adaptations required students to write answers within 2 min to 50 problems (25 on each page) of basic facts and algorithms. The basic facts and algorithms were problems matched to grade level (for Grades 1–5). The sample Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 for this task were .91, .97, and .98, respectively.

### Predictor Variables

#### Phonological Knowledge Measures

*Pseudoword reading tasks.* The Pseudoword subtest was administered from the Test of Word Reading Efficiency (TOWRE; Wagner & Torgesen, 1999). The subtest required oral reading of a list of 120 pseudowords of increasing difficulty. The dependent measure was the number of words read correctly in 45 s. The sample Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 for this task were .94, .93, and .93, respectively.

*Phonological deletion.* The Elision subtest from the CTOPP (Wagner et al., 2000) was administered. The Elision subtest measures the ability to parse and synthesize phonemes. The child was asked to say a word and then say it again with a deleted part (e.g., "Say *popcorn*. Now say *popcorn* without saying *corn*"). The dependent measure was the number of items said correctly. The sample Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 for this task were .91, .88, and .89, respectively.

#### Reading

*Word recognition.* Word recognition was assessed by the Reading subtest of the WRAT-III (Wilkinson, 1993). The task provided a list of words of increasing difficulty. The dependent measure was the number of words read correctly. The sample Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 for this task were .91, .88, and .89, respectively.

*Real word reading efficiency tasks.* The Real Word Reading subtest was administered from the TOWRE (Wagner & Torgesen, 1999). The subtest required oral reading of a list of 120 real words of increasing difficulty. The dependent measure was the number of words read correctly in 45 s. The sample Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 for this task were .88, .96, and .93, respectively.

*Reading comprehension.* Reading comprehension was assessed by the Passage Comprehension subtest from the Woodcock Reading Mastery Test—Revised (WRMT-R; Woodcock, 1998). The dependent measure was the number of questions answered correctly. The sample Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 for this task were .93, .87, .84, respectively.

### *Naming Speed*

*Letter naming speed.* The administration procedures followed those specified in CTOPP. Participants were required to name the letters as quickly as possible for each of two stimulus arrays containing 36 letters, for a total of 72 letters. The dependent measure was the total time to name both arrays of letters. The correlation between arrays in Forms A and B was .90. The sample Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 for this task were .91, .87, and .81, respectively.

### *STM Measures*

Four measures of STM were administered: Forward and Backward Digit Span, Word Span, and Pseudoword Span. The digit subtest from the WISC-III was administered. The Forward Digit Span task required participants to recall and repeat in order sets of digits that were spoken by the examiner and that increased in number. The Backward Digit Span task from the WISC-III required participants to recall sets of digits in reverse order and was administered in the same manner as the Forward Digit Span task. The dependent measure was the highest set of items recalled in order (range = 0 to 6 for Backward Digit Span).<sup>1</sup> The Word Span and Pseudoword Span tasks were presented in the same manner as the Forward Digit Span measure. The Word Span task was previously used by Swanson, Ashbaker, and Lee (1996). The word stimuli are one- or two-syllable high-frequency words. Students are read lists of common but unrelated nouns and then are asked to recall the words. Word lists gradually increased in set size, from a minimum of two words to a maximum of eight. The Phonetic Memory task (Pseudoword Span task; Swanson & Berninger, 1995) uses strings of nonsense words (one syllable long), which are presented one at a time in sets of 2–6 nonwords. The dependent measure for all STM measures was the largest set of items retrieved in the correct serial order (range = 0–7). The Cronbach's alpha coefficients on scores for Waves 1, 2, and 3, respectively, scores for the Forward Digit Span were .78, .77, and .84; for the Backward Digit Span were .46, .62, and .48; for the Word Span were .70, .77, and .75; and for the Pseudoword Span were .58, .73, and .69.

### *WM Measures*

The WM tasks in this study required children to hold increasingly complex information in memory while responding to a question about the task. The questions served as distracters to item recall because they reflected the recognition of targeted and closely related nontargeted items. A question was asked for each set of items and the tasks were discontinued if the question was answered incorrectly or if all items within a set could not be remembered. For this study, WM tasks were divided into those identified in the literature as tapping executive processing (Listen-

ing Span, Updating, Sentence–Digit task, Semantic Association task) and those assumed to tap the visual–spatial system (Visual–Matrix task, Mapping–Direction task). The complete description of the administration and scoring of the tasks is reported in Swanson (1995). Task descriptions follow.

### *Executive Processing*

*Listening sentence span.* The children's adaptation (Swanson, 1992) of Daneman and Carpenter's (1980) Sentence Span Task was administered. This task required the presentation of groups of sentences, read aloud, for which children tried to simultaneously understand the sentence contents and to remember the last word of each sentence. The number of sentences in the group gradually increased from two to six. After each group was presented, the participant answered a question about a sentence and then was asked to recall the last word of each sentence. The dependent measure was the total number of correctly recalled word items up to the largest set of items (e.g., Set 1 contained two items, Set 2 contained three items, Set 3 contained four items, etc.) in which the process question was also answered correctly. Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 were .89, .89, and .83, respectively.

*Semantic association task.* The purpose of this task was to assess the participant's ability to organize sequences of words into abstract categories (Swanson, 1992, 1995). The participant was presented a set of words (one every 2 s per word), asked a discrimination question, and then asked to recall the words that "go together." For example, a child was first presented a list of the following words, *yellow, cake, red, cookies*, then was asked a discrimination question, "Did I say *red* or *orange*?" and then was asked to group the words by category (e.g., colors and food). In other words, the task required participants to transform information encoded serially into categories during the retrieval phase. The range of set difficulty was two categories with two words in each category in Set 1 and five categories of four words each in Set 7. Thus, the number of words in each set varied from 4 words in Set 1 to 20 words in Set 7. The dependent measure was the total number of correctly recalled words in which all items in the set were recalled and the process question for the set of words was answered correctly. Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 were .84, .91, and .93, respectively.

*Digit/sentence span.* This task assesses the child's ability to remember numerical information embedded in a short sentence (Swanson, 1992, 1995). For example, Item 3 states, "Suppose

<sup>1</sup> There is some debate as to whether the backward Digit Span test better captures WM than STM (Gathercole, Pickering, Ambridge, & Wearing, 2004). As suggested by Colom, Abad, et al. (2005) and Engle et al. (1999), the numbers reversed task is assumed to be mainly a short-term processing capacity measure. Colom, Flores-Mendoza, et al. (2005) found forward and backward span measures to be similar measures of STM storage (see p. 1010 for discussion). Further, as stated by Engle et al. (1999) "Rosen and Engle (1997) showed that the backward and forward word task displayed similar effects of phonological similarity ... suggesting that a simple transposition of order would be insufficient to move a task from the STM category to the WM category" (p. 314). In addition, Swanson, Mink, and Bocian (1999) found that with young children at risk for learning problems that both forward and backward digits loaded on the same factor as phonological processing (see Table 5).



somebody wanted to have you take them to the supermarket at 8-6-5-1 Elm Street." The numbers are presented at 2-s intervals, followed by a process question ("What was the name of the street?"). The dependent measure was the total number of correctly recalled digits, in which all items in the set were recalled (set size ranged from 2 to 14 digits) and the process question for the set of digits was answered correctly. Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 were .88, .91, and .93, respectively.

*Updating.* This experimental updating task was adapted from Morris and Jones (1990). For this task, a series of one-digit numbers was presented that varied in set lengths of nine, seven, five, or three. No digit appeared twice in the same set. The examiner stated that the list may be either long or short and the participant should only remember the last three numbers in the same order as presented. Each digit was presented at approximately 1-s intervals. The four practice trials used list lengths of three, five, seven, and nine digits each, in random order. It was stressed that some of the lists of digits would be short, so they should not ignore any items. That is, to recall the last three digits in an unknown ( $N = 3, 5, 7, 9$ ) series of digits, one must keep available the order of old information (previously presented digits), along with the order of newly presented digits. The dependent measure was the total number of lists correctly repeated (range = 0–16). Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 were .93, .89, and .89, respectively.

### *Visual-Spatial Sketchpad*

*Visual matrix task.* The purpose of this task was to assess the ability of participants to remember visual sequences within a matrix (Swanson, 1992, 1995). Participants were presented a series of dots in a matrix and were allowed 5 s to study the matrix. The matrix was then removed and participants were asked, "Are there any dots in the first column?" After answering the discriminating question (by circling *y* for yes or *n* for no), students were asked to draw the dots they remembered seeing in the corresponding boxes of their blank matrix response booklets. The task difficulty ranged from a matrix of 4 squares and 2 dots (Set 1) to a matrix of 45 squares and 12 dots (Set 11). The dependent measure was the total number of items correct up to the highest set, in which the process question was answered correctly. Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 were .78, .79, and .84, respectively.

*Mapping and directions.* This task required a child to remember a sequence of directions on a map (Swanson, 1992, 1995). The experimenter presented a street map with dots (stop lights) connected by lines and arrows, which illustrated the direction a bicycle would go to follow a route through the city. After the map was removed, the child was asked a process question ("Were there any stop lights [dots] on the first street [column]?"). The child was then presented a blank matrix, on which to draw the street directions (lines and arrows) and stop lights (dots). The task difficulty on this subtest ranged from 2 dots and 3 lines (Set 1) to 20 dots and 23 lines (Set 9). The dependent measure was the total number of items correct up to the highest set, in which the process question was answered correctly. Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 were .78, .79, and .84, respectively.

### *Fluency and Inhibition Measures*

Two tasks that we assume capture different aspects of controlled attention are fluency and random generation. Both tasks measure inhibition but emphasize different aspects. The fluency task requires individuals to spontaneously generate words in response to a category cue (e.g., generate animal names) or specific letter cue (generate words that begin with the letter *B*). These tasks have been associated with the executive processing that involves the controlled search-for words (e.g., see Rende, Ramsberger, & Miyake, 2002, for review). That is, participants are directed to activate needed information (animal names) while controlling the repetition of exemplars.

In the random generation procedure, on the other hand, participants are asked to keep track of the number of times that the items have been generated and to inhibit well-known sequences such as 1, 2, 3, 4 or a, b, c, d. This task differs somewhat from the fluency measure because participants must suppress rote or habitual responses (saying the letters of the alphabet in order) in order to quickly complete the task. Thus, during random generation the central executive system acts as a rate-limited filtering device that filters out habitual responses (see Towse, 1998, for a review of this measure).

*Categorical fluency.* This experimental measure was adapted from Harrison, Buxton, Husain, and Wise (2000). Children were given 60 s to generate as many names of animals as possible. Children were told, "I want to see how many animals you can name in a minute." The dependent measure was the number of different words correctly stated within 60 s. Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 were .76, .76, and .75, respectively.

*Letter fluency.* This experimental measure was adapted from Harrison et al. (2000). Children were given 60 s to generate as many words as possible beginning with the letter *B*. The dependent measure was the number of words correctly stated in 60 s. Cronbach's alpha coefficients on scores for Waves 1, 2, and 3 were .76, .76, and .75, respectively.

*Random generation of letters and numbers.* Each child was asked to write as quickly as possible numbers (or letters) first in sequential order to establish a baseline. Children were then asked to quickly write numbers (or letters) in a random nonsystematic order. Scoring included an index for randomness, information redundancy, and percentage of paired responses to assess the tendency of participants to suppress response repetitions. The measure of inhibition was calculated as the number of sequential letters or numbers divided by number of correctly unordered numbers or letters. It was our assumption that scores close to 1.0 would reflect a high and efficient ability to inhibit well-learned sequences. Cronbach's alpha for the Letter Generation task for Waves 1, 2, and 3 scores were .52, .69, and .63, respectively, and for the Number Generation task were .75, .68, and .69, respectively.

### *Procedure*

Five doctoral-level graduate students trained in test administration tested all participants in their schools in Waves 1, 2, and 3. One session of approximately 45–60 min was required for small group test administration and one session of 45–60 min for indi-



vidual administration for each wave. During the group testing session, data were obtained from the Raven Colored Progressive Matrices test, WIAT, WRAT-III, mathematical word problem solving process booklets, visual matrix test, and arithmetic calculation fluency. The remaining tasks were administered individually. We counterbalanced test administration to control for order effects. Task order was random across participants within each test administrator.

### *Statistical Method*

*Sequence of analysis.* The results are organized into five parts. First, we determined whether the memory tasks fit a three-factor model (central executive, phonological loop, visual-spatial sketchpad). Thus, we computed a confirmatory factor model to determine the adequacy of fit to the data. In addition, we converted tasks that were conceptually related to single latent variables to reduce the sample-task ratio and to simplify the analyses.

Second, we compared performance of children at risk and not at risk for SMD across the testing waves. Because some initial differences emerged between the groups on measures of the Raven Colored Progressive Matrices test, scores from this test were used as a covariate in the multivariate analysis of covariance (MANCOVA). This analysis addressed the question as to whether children identified at risk or not at risk for SMD in Wave 1 vary on measures of growth in problem solving and WM across the three testing waves.

Third, we tested for convergence across the three age cohorts. Convergence consists of testing whether cohort groups tested at the same age overlap in performance and the parameters of growth are equal for each of the cohort groups (see E. R. Anderson, 1993, for a review). This test was necessary because we were using a cohort-sequential design. The design assumes that the parameters of growth (levels, slopes, and error terms) are invariant across the three age groups. Therefore, it was necessary to establish convergence across the three cohorts (children who started testing at Grade 1 vs. children who started testing at Grade 2 vs. children who started testing at Grade 3). To accomplish this, we conducted structural equation modeling using EQS 6.0 (Bentler, 2005) to determine whether the latent measures were invariant across the samples.

Fourth, after factor invariance was tested across the latent measures, a growth curve analysis was conducted using hierarchical linear modeling (HLM; Bryk & Raudenbush, 1992; Singer, 2002). We sought to isolate the components of WM that best related to word problem solving. We report our procedures for using HLM modeling in the next section. Overall, this analysis addressed the question as to whether growth in WM was related to growth in word problem solving. We also addressed the question as to whether ability groups varied in their rate of growth. The HLM method overcomes some of the limitation of our MANOVA because it does not assume that an equal number of repeated observations are taken for each individual or that all individuals were measured at the same time point. The model also allowed us to use random effects to model the continuous functions of age. Further, the HLM procedure does not require that missing data be ignored and provides a valid means to addressing standard errors. In contrast to traditional MANOVA repeated measures in which significance is tested against the residual error, the test of fixed

effects in mixed models is tested against the appropriate error terms as determined by the model specification.

The final section of the results focused on Wave 1 predictions of Wave 3 word problem solving. In this analysis, we were not interested in growth per se or the multilevel structure of the data but rather in how much of the variability in problem solving in Wave 3 was accounted for in Wave 1 data. Because the explained proportion of variance is somewhat problematic in HLM analysis (e.g., a large number of variance components, negative  $R^2$  are possible, etc.; for discussion, see Snijders & Bosker, 1999, pp. 99–100), we used hierarchical regression models to simply isolate unique processes in Wave 1 that underlie word problem solving performance in Wave 3. This analysis allowed us to address the questions related to the three models (knowledge base, phonological system, central executive system) provided in the introduction. The key question addressed was whether performance related to three components of WM was related to predictions of problem solving accuracy and whether those predictions were mediated by individual differences in knowledge base, phonological processing, and/or executive processing.

A key assumption of this study is that partialing out the influence of STM from WM leaves residual variance related to controlled attention (Engle et al., 1999). Several studies have shown that WM and STM are distinct but highly related processes (e.g., see Heitz, Unsworth, & Engle, 2005, for a review). For example, Engle et al. (1999) investigated the relationship among measures of STM, WM, and fluid intelligence in adults. They found that STM and WM tasks loaded on two different factors. Although strong correlations emerged between the two factors ( $r = .70$ ), they found that a two-factor model fit the data better than a one-factor model. Further, they found that by statistically controlling the variance between WM and STM factors, the residual variance related to the WM factor was significantly correlated with measures of intelligence. That is, they found a strong link between the latent measures of WM, but not STM, to fluid intelligence (e.g., also see Conway, Cowan, Bunting, Theriault, & Minkoff, 2002). They interpreted their findings as suggesting that the residual variance related to the WM factor corresponded to controlled attention of the central executive system. Thus, in our regression analysis, we entered measures of STM and WM simultaneously into the regression model to determine whether unique variance related to WM would predict problem solving accuracy.

*HLM.* As mentioned previously, we examined growth in problem solving using a multilevel framework referred to as HLM. We applied growth modeling to the study of intraindividual change in problem solving and WM over the 3-year period via the PROC MIXED program in SAS 9.1 (SAS Institute, 2003). The HLM procedure allowed us to determine both the average rate of change and individual variability in change over time. Age was the variable that represented the passage of time in our growth model. Because we had only three data points at best, our focus was on linear change and, therefore, a curvilinear relationship could not be reliably calculated. To interpret the results, we centered age at 9.7 (mean age at Wave 3), so that intercepts reflected the expected performance at that age. (It is important to note that slope remains the same whether the mean age is centered at Wave 1, 2, or 3. However, the correlation between the intercept and slope does vary as a function of different centering.) Thus, we centered our data on



the mean age at Wave 3 because we were interested in the final status of individual children after 3 years of math instruction.

Our growth model yielded parameter estimates that defined both the overall trajectory of the sample (fixed effects) and deviations in the overall trajectories (random effects). The model is expressed as:

$$y_{ij} = \beta_0 + \beta_1(\text{age}_{ij}) + U_{0j} + U_{1j}(\text{age}_{ij}) + R_{ij},$$

where  $y_{ij}$  is the dependent variable (problem solving) measured at time  $i$  in child  $j$ ;  $\text{age}_{ij}$  is the child  $j$ 's age at time  $i$ ;  $\beta_0$  is the average intercept at 9.7;  $U_{0j}$  is the random intercept for child  $j$ ;  $U_{1j}$  is the random age slope for child  $j$ ; and  $R_{ij}$  is the residual for child  $j$  at time  $i$ . The between-children variance components,  $\tau^2_0 = \text{Var}(U_{0j})$  and  $\tau^2_1 = \text{Var}(U_{1j})$ , reflect individual differences in level (intercept) and rate of change (slope), respectively. Thus, we estimated the association between the outcome (problem solving) and repeated measures of age across the 3-year time periods. We refer to this as our *unconditional growth model*.

Clearly, performance on word problems will be influenced by instructional procedures in the math classroom. At the start of the study, children were nested within different classrooms that reflect different instructional approaches of the math teacher (for the majority of participants, math instruction was in the homeroom class). Our analyses of random effects for intercept and slope included children nested within classroom (or math teacher). Thus, we modified the random effects (between-children effects) as  $\tau^2_0 = \text{Var}(U_{0j}) \times (\text{Teacher or Math Class})$  and  $\tau^2_1 = \text{Var}(U_{1j}) \times (\text{Teacher or Class})$  to reflect individual differences in level (intercept) and rate of change (slope), respectively. Classroom (math teacher) placement was not a continuous variable across the testing waves. For our analysis we selected Wave 1 math teachers for the nesting effect on the basis of our results from our administration of a National Science Foundation (2000) National Survey of Science and Mathematics Education. This survey showed the greatest amount of variance in instructional approaches among teachers occurred during Wave 1 when compared with other testing waves. Thus, our random effects for intercept and slope were nested within classroom at Year 1 (thus, we had randomly varying intercepts and growth rates within classroom). It is also important to note that we initially used a three-level model (Level 1 effects within person, Level 2 effects between persons nested within classroom, Level 3 effects between classrooms; see Singer, 2002, p. 167 for an example of this three-level growth model). However, because the random effects at Level 3 were not significant and less than 2% of the variance, these random effects were dropped from the analysis. Our approach was consistent with the model specification steps outlined in Snijders and Bosker's study (1999; see pp. 94–97), which called for excluding nonsignificant random effects.

In summary, our HLM (Level 1 within person, Level 2 between persons nested within classroom) included random effects from intercepts and slopes for children nested within the classroom of the math teacher. We used this unconditional growth model to examine problem solving performance as a function of the intercept and slope. For the fixed effect, the intercept provided information on the average level of the dependent variable at age 9.7 and the average rate of change across individuals. For the random effects, the intercept represented the variation (variance) around the intercept and the slope indicated whether there was variance related to change overtime. Significant random effects indicated that children differed in the level (intercepts) and/or rate of change (slopes).

After establishing our unconditional growth model, we tested whether entering the components of WM into the model explained any statistically significant associations obtained related to fixed effects and random effects. This is referred to as a *conditional model*. When one or more predictors are introduced into the conditional model, the reductions in the magnitude of the various components when compared with the unconditional model are analogous to effect sizes (Snijders & Bosker, 2003). This is similar to the use of  $R^2$  in linear regression models. The primary distinction between a linear regression and HLM is that several  $R^2$  values are relevant to HLM because there are several variance components.

Reliability estimates for the HLM model are based on the random effects. The random effects of the unconditional model represent the proportion of variance in that effect (i.e., parameter specific rather than error variance). If the random effect is significant and has satisfactory reliability, then it is appropriate to test whether additional variables can explain some of the variance in the unconditional model. Specifically, the conditional model predicted that the level and slope of problem solving performance would be associated with WM performance. To evaluate the compatibility of the data with our conditional model, we tested the significance of the model change. This was done by comparing the differences between the deviance values (i.e., the likelihood value for the correspondence between model and data) from the unconditional and conditional growth model. These are chi-square values, and the number of parameters added for the conditional model serves as degrees of freedom. In general, models with lower deviance fit better than models with higher deviance values. The deviance test can be used to perform a formal chi-square test in order to compare a model that adds certain effects versus a model that excludes them. For the present study, a significantly lower deviance score for the conditional model indicated that the conditional model showed a better fit to the data than the unconditional model.

We also compared children at risk for SMD and not at risk on parameter estimates of the intercept and slope. Because our classification of risk for SMD occurred at Wave 1, we centered age at Wave 3. Ability group differences have been established at Wave 1 (Swanson & Beebe-Frankenberger, 2004) and therefore it would be more informative about the stability of ability group differences by centering performance at Wave 3. Within this hierarchical model, we conducted several analyses comparing the risk and nonrisk groups identified in Wave 1 on intercept and growth scores across an array of measures.

**Missingness.** Missing data occur in longitudinal designs whenever an intended measurement is not obtained. We experienced some loss in our sample in Year 3 ( $N = 353$  in Wave 1 and  $N = 302$  in Wave 3). HLM allows for an incomplete data set, using data from all participants with at least two points. We used maximum likelihood procedures to determine parameter estimates because the maximum likelihood estimation procedure has several advantages over other missing data techniques (see Peugh & Enders, 2004, for discussion).

## Results

The means and standard deviations for measures of intelligence, accuracy in recognizing problem solving components, problem solving solution accuracy, arithmetic calculation, phonological processing, reading, STM, WM, inhibition, and fluency for Waves 1, 2, and 3 are shown in Appendix A (starting with Wave 1 at



Grade 1), B (starting with Wave 1 at Grade 2), and C (starting with Wave 1 at Grade 3). Prior to testing our hypothesis, we considered normality of the data. Measures meet standard criteria for univariate normality with skewness for all measures less than 3 and kurtosis less than 4. Univariate outliers were defined as cases more than 3.5 standard deviations from the means. We examined multivariate outliers by calculating Mahalanobis'  $d^2$ . None of the cases were deemed outliers.

### *Confirmatory Factor Analysis and Data Reduction*

A critical assumption of this study was that three components were represented in our WM battery: executive processing, phonological loop, and visual-spatial sketchpad. In order to test this assumption, we ran a confirmatory factor analysis for Wave 1 data using the CALIS (Covariance Analysis and Linear Structural Equation) software program (SAS Institute, 2003) with the four WM tasks (Listening Span, Semantic Association, Digit-Sentence Span, Updating) loading on one factor, the four STM tasks (Forward Digit Span, Backward Digit Span, Pseudoword Span, Word Span) loading on a second factor, and the two visual WM tasks loading on a third factor (Visual Matrix, Mapping-Directions). The fit statistics were .91 for the comparative fit index (CFI; Bentler & Wu, 1995) and .05 for the root-mean-square residual error of approximation (Jöreskog & Sörbom, 1984), indicating a good fit to the data. All standardized parameters were significant at the .01 level. As we had no theoretical reasons for relying on a one-factor or two-factor model, the three-factor model was accepted as a good fit to the data for the total sample in Wave 1.

Because of the large number of variables, it was necessary to reduce two or more measures that were conceptually related to latent measures (factor scores). The weighting of these variables for Waves 1 and 2 are reported in Swanson's (2006) study. Task weightings for the latent measures used in regression analysis (to be discussed) are reported in Table 7. We used the SAS CALIS program to create factor scores for each set of measures with two or more variables. This procedure allowed us to calculate for each testing wave standardized beta weights. On the basis of the standardized loadings, we computed factor scores by multiplying the  $z$  score of the target variable by the factor loading weights for the total sample (see Nunnally & Bernstein, 1994, p. 518 for calculation procedures).<sup>2</sup> Factor scores were created for problem solving accuracy (WISC-III: Arithmetic; problem solving semantic structure varied), math calculation (WRAT: Math; WIAT: Math, Computation Fluency), word problem solving components (correctly identifying the word problem propositions-related questions, numerical information, goals, arithmetical operations, algorithm, and irrelevant information), reading (WRAT: Reading; WRMT: Reading Comprehension), phonological processing knowledge (TOWRE: Pseudowords, Elision), inhibition (random generation of letters and random generation of numbers), fluency (Phonological and Semantic Fluency), speed (rate naming speed of numbers and letters), phonological loop (Forward Digit, Backward Digit, Pseudoword Span, Real Word Span), visual-spatial sketchpad (Visual Matrix, Mapping/Directions), and executive processing (Listening Sentence Span, Digit/Sentence, Semantic Association, Updating).

### *Ability Group and Wave Comparisons*

Figure 1 shows the latent measures as mean  $z$  scores for each measure used in the subsequent analysis as a function of the three

cohorts and two ability groups across the three testing waves. The figure shows the starting points by age for three cohorts at Wave 1 and follows their performance across the three testing waves. The age range was from 6 to 10 (Grades 1–5). Figure 1 also shows children identified at risk for SMD and those not at risk within each cohort. All figures were scaled to show a range of 2.5 standard deviations above and below a mean  $z$  score of 0 based on Wave 1 performance.

As shown in Figure 1, measures that showed the greatest increase from ages 6 to 10 (Grades 1–5) were math calculation, reading, and random generation. Measures showing the smallest rate of change ( $< .25 SD$ ) were performance measures of fluency, phonological loop, and knowledge of word problem solving components. Across all measures, children identified at risk were lower in performance than those not at risk. It is important to note that problem solving from the WISC-III Arithmetic subtest and rapid naming of numbers from the CTOPP were used as classification variables and, therefore, it would be expected that differences emerged between these two groups in Wave 1 for problem solving and speed. However, we considered it an empirical question as to whether such differences at Wave 1 would be maintained in Wave 3.

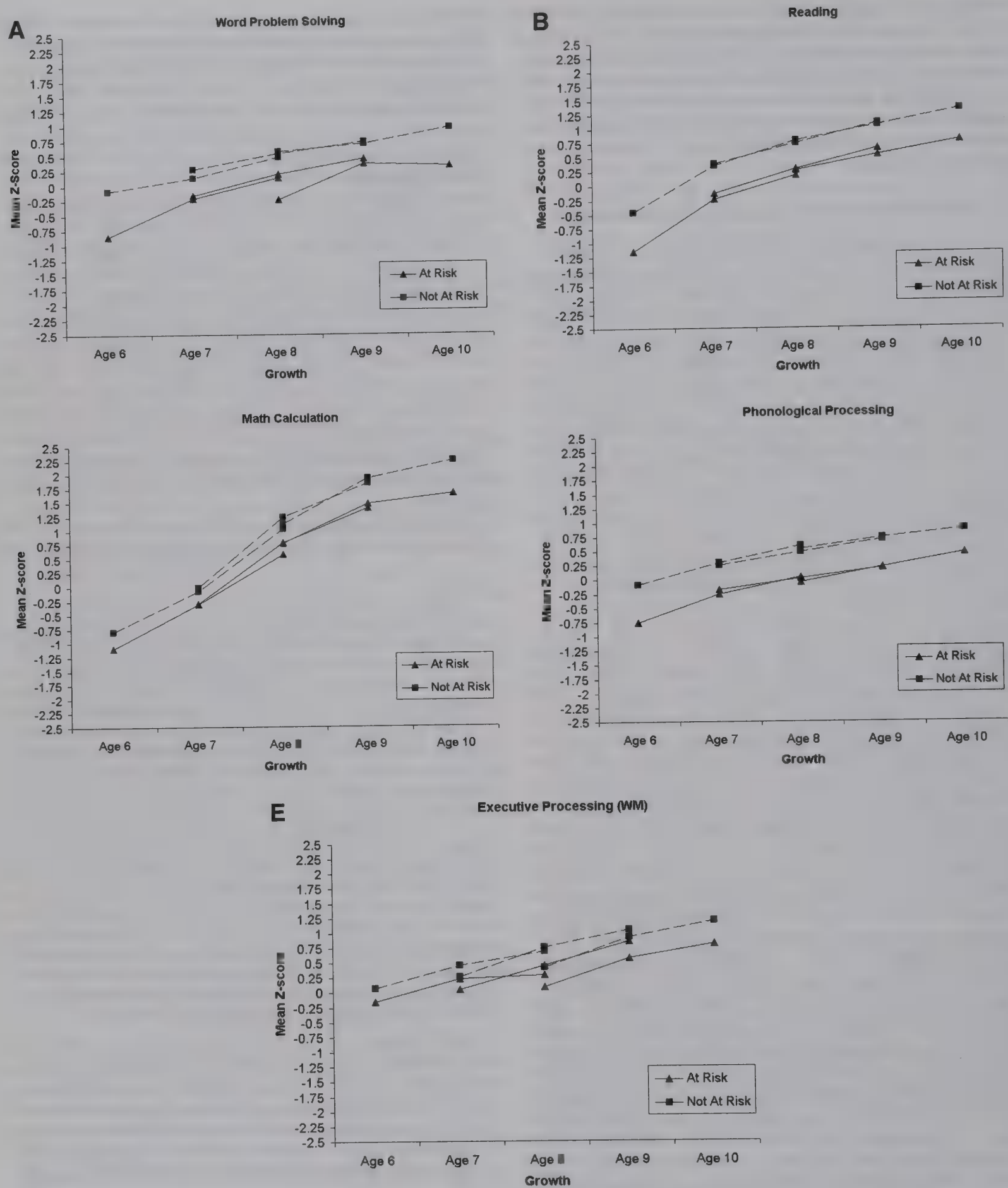
The primary question addressed in this part of the analysis was whether performance of children at risk for SMD varied from

<sup>2</sup> Consistent with other investigations of growth (e.g., Wilson et al., 2002), we converted raw scores to  $z$  scores. All measures were scaled to have a mean of 0 and standard deviation of 1 at Wave 1. Wave 2 and 3 measures were  $z$  scored on the basis of the means and standard deviations of Wave 1. It was necessary to scale to  $z$  scores across the total sample so that all parameters were on the same metric, enabling meaningful comparisons for both age and time (Waves 1, 2, and 3) (see McGraw & Jöreskog, 1971, for a discussion). In addition, because of the number of variables, we created factor scores. These scores were calculated for each of the three testing waves. This was done for measurement purposes (different variables have different weightings on a construct) and also for practical reasons: Some constructs (e.g., WM, STM) included several tasks. One reviewer questioned whether our scaling of items would best be served using an item response theory model. The item response theory framework posits a log linear, rather than a linear, model to describe the relationship between observed item responses and the level of the underlying latent trait,  $\theta$ . We agree with the reviewer's suggestion. However, as indicated by Nunnally and Bernstein (1994), for item response theory models "very large normative bases are required to implement all but the simplest and, therefore, sometimes unrealistic models using current estimation algorithms" (p. 396). In addition, as stated by Plewis (1996),

The feasibility of using vertical equating in the context of growth studies spanning a narrow range with different tests used at different ages has yet to be demonstrated. Certainly it is difficult to see how item response theory, which depends on fixing the mean and standard deviation of the ability distribution in order to estimate both item parameters . . . and an individual's ability, could be adapted to deal with growth without making arbitrary scaling assumptions. (p. 28)

Because we cannot address these issues with our current sample size (e.g., our sample is not large enough to establish anchor points), we attempted to place our values on the same measurement scale across the groups and investigate linear change. Converting data to  $z$  scores across the sample is appropriate when testing covariance structures (e.g., see Reise, Widaman, & Pugh, 1993, p. 557, for an example), as we have done in this study.





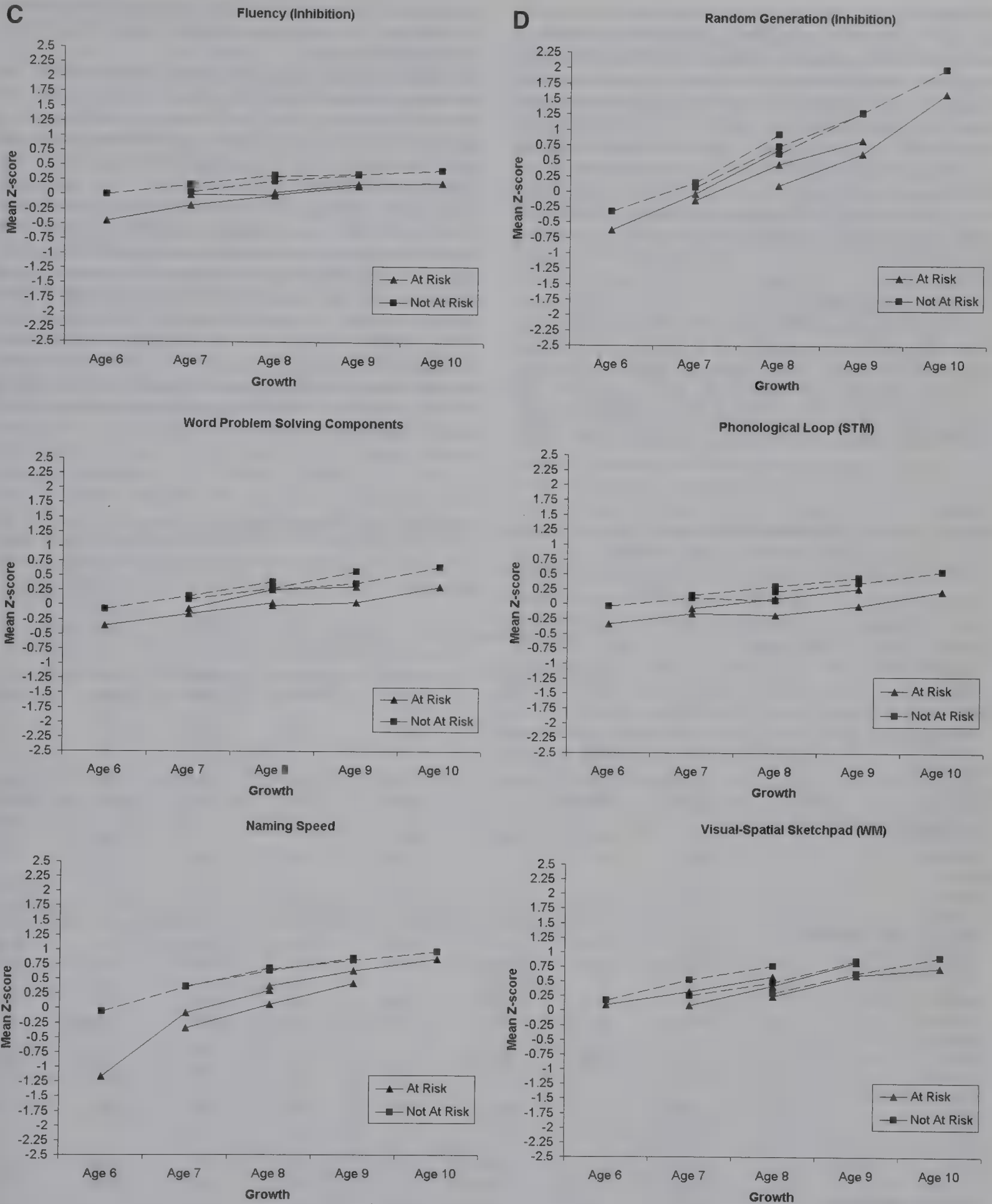


Figure 1 (opposite). Mean z scores for achievement and cognitive domains across testing waves as a function of age and ability group. STM = short-term memory; WM = working memory.



children not at risk on latent measures across testing waves. To address this question, we computed a 2 (risk group: children not at risk vs. SMD)  $\times$  11 (number of domains: problem solving accuracy, math calculation, word problem solving components, reading, phonological processing knowledge, inhibition, fluency, speed, phonological loop, visual-spatial sketchpad, and executive processing)  $\times$  3 (testing Waves 1, 2, and 3) repeated measures analysis of covariance (ANCOVA), with repeated measures on the last two factors. Although Raven scores were in the normal range for the at-risk group, an advantage was found for the not-at-risk group. Thus, Raven scores served as a covariate in the analysis. The covariate (Raven scores) was significant,  $F(1, 266) = 10.01$ ,  $MSE = 4.34$ ,  $p < .001$ , but the Ability Group  $\times$  Covariate interaction was not significant,  $F(1, 266) = 1.59$ ,  $p = .20$ , meeting the assumptions of a MANCOVA (homogeneity of the slopes). Thus, the covariate was judged to be reliable for the covariance analysis. Because there was a violation of sphericity (Mauchly's criterion reported probability  $< .0001$ ), we used the Greenhouse-Geisser and the Huynh-Feldt probability values. Because the samples of children not at risk and at risk for SMD

were matched on age, age was not considered as a covariate in this analysis. (The influence of age will be considered in the subsequent analyses.)

Omnibus  $F$  effects associated with the risk status, wave, and domain are reported below. The mean  $z$  scores of the latent measures are shown in Table 1 for Waves 1, 2, and 3. As shown in Table 1, the general pattern in the results was that children not at risk scored higher than those at risk for SMD, and scores were higher in testing Waves 2 and 3 than in Wave 1. Also provided in Table 1 is the difference in  $z$  scores between Wave 3 and Wave 1. As shown in Table 1, mean difference scores (Wave 3 minus Wave 1) greater than 1.0 standard deviation occurred for the measures of math calculation and inhibition. Although it is important to note that not all data can be used in this analysis (missing cells across waves are dropped), the results do show that differences between Wave 1 and Wave 3 were larger ( $z$  score differences  $> .40$ ) for the children not at risk on measures of math calculation and knowledge of problem solving components. In contrast, difference scores were larger for the at-risk group on measures of problem solving accuracy and naming speed.

Table 1  
Mean  $z$  Scores for Each Domain (Composite Score) as a Function of at Risk for Serious Math Difficulties (SMD;  $n=104$ ) and Not at Risk for Serious Math Difficulties (NSMD;  $n=166$ )

Domain	Wave 1		Wave 2		Wave 3		Differences	
	$M$	$SD$	$M$	$SD$	$M$	$SD$	LSM	$SD$
Problem solving								
SMD	-0.50	0.60	0.03	0.41	0.28	0.39	0.82	0.55
NSMD	0.32	0.37	0.50	0.40	0.72	0.48	0.40	0.45
Math								
SMD	-0.36	0.62	0.42	1.09	0.98	0.96	1.58	1.04
NSMD	0.21	0.70	1.30	1.41	1.78	0.92	2.12	1.06
Reading								
SMD	-0.57	0.84	0.07	0.70	0.43	0.68	1.04	0.50
NSMD	0.37	0.67	0.78	0.50	1.10	0.48	0.73	0.41
Phonological knowledge								
SMD	-0.45	0.62	-0.07	0.57	0.18	0.59	0.63	0.44
NSMD	0.30	0.63	0.54	0.54	0.74	0.50	0.44	0.45
Fluency								
SMD	-0.22	0.53	-0.02	0.44	0.09	0.38	0.32	0.44
NSMD	0.14	0.52	0.28	0.43	0.35	0.44	0.21	0.55
Problem solving components <sup>a</sup>								
SMD	-0.23	0.40	0.12	0.52	0.55	0.63	0.76	0.56
NSMD	0.13	0.34	0.59	0.53	1.17	0.61	1.05	0.52
Speed								
SMD	-0.67	1.17	0.08	0.64	0.44	0.53	1.12	0.84
NSMD	0.40	0.51	0.65	0.42	0.85	0.40	0.44	0.35
Inhibition								
SMD	-0.35	0.60	0.21	0.71	0.87	0.71	1.18	0.70
NSMD	0.26	0.73	0.83	0.85	1.54	0.99	1.34	0.72
Phonological loop								
SMD	-0.22	0.38	-0.17	0.40	-0.02	0.41	0.23	0.38
NSMD	0.12	0.39	0.15	0.44	0.38	0.47	0.25	0.44
Visual-spatial sketchpad								
SMD	-0.13	0.49	-0.08	0.60	0.24	0.72	0.39	0.78
NSMD	0.05	0.56	0.18	0.81	0.62	0.90	0.56	0.93
Executive working memory								
SMD	-0.19	0.39	0.06	0.55	0.37	0.61	0.59	0.57
NSMD	0.09	0.48	0.56	0.61	0.90	0.73	0.79	0.73

Note. LSM = least square means partialled for Raven scores.

<sup>a</sup> Knowledge of problem solving components.

The repeated measures ANCOVA indicated that significant main effects emerged for ability group,  $\lambda = .50$ ,  $F(33, 235) = 6.92$ ,  $p < .0001$ ,  $\eta^2 = .50$ ; domain,  $\lambda = .77$ ,  $F(10, 258) = 7.62$ ,  $p < .0001$ ,  $\eta^2 = .23$ ; and testing wave,  $\lambda = .83$ ,  $F(2, 266) = 27.08$ ,  $p < .0001$ ,  $\eta^2 = .17$ . Significant interactions emerged for the Group  $\times$  Domain interaction,  $\lambda = .77$ ,  $F(11, 259) = 7.61$ ,  $p < .0001$ ,  $\eta^2 = .23$ ; the Domain  $\times$  Wave interaction,  $\lambda = .73$ ,  $F(11, 248) = 4.41$ ,  $p < .0001$ ,  $\eta^2 = .27$ ; the Ability Group  $\times$  Wave interaction,  $\lambda = .97$ ,  $F(2, 266) = 3.24$ ,  $p < .05$ ,  $\eta^2 = .03$ ; and the Ability Group  $\times$  Domain  $\times$  Wave interaction,  $\lambda = .66$ ,  $F(11, 248) = 6.19$ ,  $p < .001$ ,  $\eta^2 = .34$ .

A test of simple effects indicated that comparisons between at-risk and not-at-risk students were significant for each wave and domain. However, greater differences emerged between some domains than others. In general, a comparison of ability groups across domain (collapsed across time) indicated that the differences in favor of children not at risk, when compared with children at risk for SMD, were greater on all measures, except for the fluency and the storage components of WM (visual-spatial sketchpad and phonological loop). Further, when the difference scores between Wave 3 and Wave 1 served as a dependent measure (Wave 3 minus Wave 1), significant differences ( $ps < .01$ ) in ANCOVAs were found between ability groups for all measures except for the measures of STM (phonological loop), fluency, visual spatial WM (visual-spatial sketchpad), and inhibition ( $ps > .05$ ). As shown in Table 1, significant differences in scores emerged in favor of the nonrisk group on the remaining measures (all  $ps < .01$ ).

In summary, the general trend was that children not at risk scored higher than those at risk, and higher difference scores emerged for academic domains (e.g., math calculation) when compared with cognitive domains (e.g., inhibition).

### Convergence

A key assumption in growth modeling using a cohort sequent design is convergence. Convergence consists of testing whether cohort groups tested at the same age overlap in performance and

the parameters of growth (the levels, slopes, and error terms) are equal for each of the cohort groups (see E. R. Anderson, 1993, for a review). To address these assumptions, we used structural equation modeling (EQS 6.0; Bentler, 2005) to test whether the underlying linear growth processes (mean and variance of the slope and intercept factors) were invariant for the three cohort age groups (children in Wave 1 who started at Grade 1, 2, or 3). In structural equation modeling, it is possible to test the invariance (equality) on all parameter estimates across multiple groups. We evaluated the invariance of various sets of parameters across the three age groups. The present analysis determined whether the intercepts and slopes across all individuals depicted a common line. A statistical test of convergence was done to determine whether the parameters (the intercepts, slopes, and error terms) were equal for each of the three cohort groups. The baseline model reflected configural invariance that tested whether similar patterns (not necessarily identical) emerged between the three groups across the three testing waves. Configural invariance requires only that the number of factors and factor loading patterns be the same across the groups (Byrne, 2006, see p. 233). A goodness-of-fit statistic related to this multigroup parameterization should be indicative of a well-fitting model. A well-fitting model has CFI at or above .90 (Bentler, 2005). As shown in Table 2, the CFIs ranged from .95 to 1.00 for the configural or baseline model, indicating that groups share a common structure and pattern across all measures. As shown in Table 2, three models tested the equality (invariance) between groups on various parameters. We tested for invariance across the three samples (a) when all parameters (intercepts, slopes, and errors) were constrained to be equal (Model 1), (b) when the intercept for each sample that included the same grade levels (e.g., Grade 3 in all three samples, Grade 2 in two samples) was constrained to be equal (Model 2, slope or errors were not constrained to be equal across groups), and (c) when the intercept and slope for the same grades in each sample were constrained to be equal (Model 3, errors are not constrained to be equal). Generally, it is argued that invariance holds if goodness-of-fit (CFI) to the model (e.g., Model 1, 2, or 3) is adequate and if there is no

Table 2  
Invariance Tests Conducted Over Multiple Age Groups for a Cohort-Sequential Design

Domain	Configural model		Model 1: All constrained		Model 2: Intercepts constrained <sup>a</sup>		Model 3: Intercepts and slopes constrained	
	$\chi^2(3)$	CFI	$\chi^2(17)$	CFI	$\chi^2(9)$	CFI	$\chi^2(13)$	CFI
Word problems	10.49	.97	65.46	.91	19.92 <sup>a</sup>	.95	54.08	.94
Arithmetic calculations	87.59	.97	250.26	.71	103.98	.89	217.16	.77
Reading	51.54	.99	220.79	.90	68.37	.98	183.50	.93
Phonological processing	8.09	1.00	70.69	.97	16.65 <sup>a</sup>	.99	58.24	.98
Fluency	1.49	1.00	31.93	.92	12.03 <sup>a</sup>	.97	24.80	.98
WP components	45.58	.95	108.64	.90	53.82 <sup>a</sup>	.93	102.44	.88
Speed	25.46	.99	119.81	.94	38.43 <sup>a</sup>	.98	105.85	.95
Inhibition	5.96	1.00	28.96 <sup>a</sup>	1.00	6.28 <sup>a</sup>	1.00	21.79 <sup>a</sup>	1.00
Phonological loop	17.03	.99	38.24 <sup>a</sup>	.98	21.32 <sup>a</sup>	1.00	32.69 <sup>a</sup>	1.00
Sketchpad	12.33	.97	23.31 <sup>a</sup>	1.00	16.26 <sup>a</sup>	1.00	21.35 <sup>a</sup>	1.00
Executive working memory	3.92	1.00	34.07	.95	15.82 <sup>a</sup>	.94	31.76	.93

Note. CFI = comparative fit index; WP = word problem solving components accuracy score.

<sup>a</sup> Converged.



statistically significant difference from that of the configural model (Byrne, 2006). Although several authors consider Model 1 unrealistic (see Byrne, 2006, pp. 244–247), we found no statistical difference between the configural model (baseline model) and Model 1 for performance related to inhibition, visual–spatial sketchpad, or the phonological loop. For example, a statistical test between the phonological loop and the configural model was not significant,  $\Delta\chi^2(df = 14 [17 \text{ for Model 1}] - 3 [\text{for the configural model}]) = 21.21 (38.24-17.03), p > .05$ . As shown in Table 2, Model 2 showed convergence (invariance) for all cognitive and word problem solving measures (all  $\Delta\chi^2s > .05$ ). This finding indicated that the intercept values for overlapping ages were statistically comparable. We were unable to find convergence, however, on measures of reading and math calculation suggesting that other variables (e.g., instruction) influenced cohort effects.

HLM

As noted above, cross-group equivalencies were excellent on all measures, except two (reading and math). Thus, some sources of invariance were unaccounted for across the cohort groups. Instead of uncovering each invariant parameter across the age groups, however, we decided to center our results on age, using HLM procedures. This is because the majority of findings related to intercepts and slopes are conditional on age (see Mehta & West, 2000), and, therefore, it was not reasonable to ignore age differences. As stated by Mehta and West (2000), conventional SEM models based on sample means and covariance structures use an “inappropriate within person scaling of the time variables yielding incorrect estimates of some of the random effect and increased complexity of interpretation of age effects” (p. 24). To correct this, for our next analysis (consistent with Mehta & West, 2000) we scaled age with respect to a common origin across all individuals to reflect deviations from a common age.

Unconditional Model

Table 3 shows our individual growth modeling to the study of intraindividual change in problem solving over the 3-year period. We estimated individual-level trajectories as well as overall sample-level trajectories. Table 3 shows an unconditional mode, and Table 4 shows a conditional model. To interpret these tables, we first consider the unconditional model for the fixed effects in Table 3. The fixed effects model (which can be viewed as a one-way random effects analysis of variance model) provided the

estimates of the intercept (centered for the mean age at Wave 3) and growth. In this case, a mean *z* score of .58 was the predicted amount of solution accuracy at age 9.7 at Wave 3. The average unit of linear growth was .28. Hence, a 9.7-year-old ended Wave 3 with a *z* score of .58 and gained .28 units per testing session. Also shown are the *t* ratios (20.71 and 19.08) indicating that the parameter estimates were significantly greater than chance.

Also presented in Table 3 for the unconditional model are the model random effects for the total sample. The random effects include the intercept, slope, and within-child variance across testing waves. It is important to note that all students were exposed to variations in math instruction and, therefore, it was necessary to consider this variable in our calculation of random effects. We used as a representative variable the growth of students nested within classroom for Wave 1. Thus, our random effects (variance for intercept, slope) represented multiple observations of children (participants) nested within classrooms at Wave 1. As shown in Table 3, significant between-subjects variance emerged for the intercept (.18) and the slope (.03) estimates. Also included in the model was the within-person residual (.09). This estimate indicated that significant within-child variance emerged for problem solving accuracy across testing sessions.

Conditional Model-Group Effects

The conditional model in Table 4 shows a comparison of the two-risk groups identified at Wave 1 on problem solving accuracy. The difference between the two models is the addition of the classification (group) variable. For children at risk for SMD, the intercept (estimated *z* score for a child at 9.7 at the end of the study) for problem solving was .32, and linear growth rate was .41. These intercept and growth estimates were significantly better than chance. Likewise, for children not at risk, the intercept for word problem solving was .75 and the linear growth rate was .20. The intercept and linear growth estimates for the not-at-risk group were also significantly better than chance. In addition we determined whether the differences in estimated intercepts and growth rates varied significantly between the two groups. Ability group differences emerged in the estimated intercept values (.32 vs. .75),  $F(1, 927) = 62.93, p < .0001$  (note *df* denominator reflects estimated missing values), and the slope values (.41 vs. .20),  $F(1, 927) = 50.39, p < .0001$ . Thus, although ability group differences would be expected at Wave 1 (because at-risk groups were separated by problem solving accuracy and rapid digit naming), the results show

Table 3  
*Growth on Word Problem Solving for the Unconditional Model*

Effect	Parameter estimate	Variance estimate	SE	<i>t</i>	<i>z</i>
Fixed					
Intercept	.58		.02	20.71***	
Growth (linear)	.28		.01	19.08***	
Random (Participants × Teachers)					
Intercept		.18	.02		8.23***
Growth (linear)		.03	.006		3.95***
Residual		.09	.007		12.23***

\*\*\* *p* < .001.

Table 4  
*Growth on Word Problem Solving for the Conditional Model*

Effect	Group				<i>F</i>
	At risk for SMD		Not at risk for SMD		
	Parameter estimate	<i>SE</i>	Parameter estimate	<i>SE</i>	
	Variance estimate	<i>SE</i>	<i>z</i>		
Fixed					
Intercept	.32***	.04	.75***	.03	62.93***
Growth (linear)	.41***	.02	.20***	.02	50.39***
Random (Participants $\times$ Teachers)					
Subject	.14	.01	7.38***		
Growth	.02	.006	2.83**		
Residual	.08	.007	12.25***		

Note. SMD = serious math difficulties.

\*\*  $p < .01$ . \*\*\*  $p < .001$ .

that these differences remained at Wave 3. In addition, significant differences emerged between the two groups in linear growth, with the rate of growth higher in the at-risk than not-at-risk group. In summary, children not at risk had higher levels of performance at Wave 3 but lower linear growth estimates than children at risk.

Also presented in Table 4 for the conditional model are the model random effects for the total sample. When compared with the unconditional model, the estimated variances between subjects for the intercept (.14) and slope (.02) were significant. The random effects (estimated intercept variance between subjects and estimated variance within subjects) in the conditional model can be compared with the random effects in the unconditional model that has no ability group effects. Thus, one way of measuring how much of the variation in word problem solving accuracy was explained by creating ability group comparisons (in the conditional model) was to compute how much of the variance was reduced after comparing the variance estimates of the conditional model with those of the unconditional model. The percentage of reduction in between-subjects variance in the intercept related to the ability group conditional model was 22%:  $(.18 - .14)/.18 = .22 \times 100$ . We interpret this value by saying that 22% of explainable variation in child intercept values in Wave 3 is a function of the at-risk classification in Wave 1. (Note: the percentage reflected the fraction of variation that was explained. This is not the same as a traditional  $R^2$  statistic.) The explainable variation in slope was 33%:  $(.03 - .02)/.03$ .

Because of the number of estimates, Table 5 summarizes the fixed effects related to intercept and growth estimates for the ability groups across all measures. As shown in Table 5, the majority of estimates were significantly better than chance. The  $F$  ratios are shown comparing the ability groups on estimates of level performance and linear growth. Two important findings emerged. First, the level of performance (intercept values) was significantly lower for children at risk for SMD than children not at risk across all measures. Second, significant differences emerged in growth rates in favor of the not-at-risk group on all measures, except for the reading and phonological knowledge measures. No significant differences emerged between the two groups for growth rates on measures of fluency and STM (phonological loop).

### Conditional Model: WM Effects

One critical question that directed this study was whether growth in WM was related to growth in math problem solving accuracy. Table 6 shows the contribution of WM growth to problem solving accuracy. In this model, we tested whether the intercepts and slopes of problem solving were significantly related to the covariates. Covariates in this case were the intercept values and growth estimates for the three WM components. As shown in Table 6, a significant covariate effect was found for level of performance on all three components of WM. The estimated coefficients in Wave 3 for the phonological loop (STM; .14), executive processing (.09), and visual-spatial sketchpad (.10) were significantly related to problem solving. With respect to estimated growth rates, a significant effect occurred for the covariate of the phonological loop (−.05) and executive processing (−.10). The significant parameter estimate related to the phonological loop (−.05) indicated that a child who differed by a score of 1.0 with respect to performance on measures of the phonological loop had a growth rate that differed by −.05. Likewise, a child who differed by a score of 1.0 with respect to the executive system had a growth rate that differed by −.10.

When compared with the unconditional model in Table 3, the conditional model in Table 6 with WM components reduced variance related to between-children intercepts by 61%  $(.18 - .07)/.18$  and slopes by 66%  $(.03 - .01)/.03$ . We determined whether this conditional model provided a good fit to the data by calculating its deviance value (i.e., lack of correspondence between model and data). The log likelihood  $-2\ln(L)$  was 1,252.3 for the unconditional model and 954.7 for the conditional growth model. A significant chi-square indicated that the conditional model showed a better fit to the data than the unconditional model,  $\chi^2(6) = 297.6, p < .001$ .

In summary, ability group effects related to the level of performance in the final wave were higher for children not at risk across all measures when compared with children at risk for SMD. Significant growth differences in favor of the not-at-risk group emerged on all measures, except reading and phonological knowledge. No significant group differences emerged in growth rates on



Table 5  
*Fixed Effects for the Conditional Model Comparing Intercept and Growth for Children at Risk for Serious Math Difficulties (SMD) and Not at Risk on Measures of Achievement and Cognition*

Measure	Group				F
	At risk for SMD		Not at risk for SMD		
	Estimate	SE	Estimate	SE	
Math calculation					
Intercept	1.09***	.07	1.77***	.05	58.10***
Growth	0.58***	.02	0.76***	.02	28.27***
Reading					
Intercept	0.68***	.05	1.09***	.04	34.27***
Growth	0.48**	.02	0.34***	.01	32.25**
Phonological knowledge					
Intercept	0.27**	.05	0.70***	.04	43.48***
Growth	0.28**	.01	0.21**	.01	9.97***
Fluency					
Intercept	0.12**	.04	0.35***	.02	20.72***
Growth	0.13***	.01	0.10***	.01	1.68
Problem solving component					
Intercept	0.54***	.04	0.97***	.03	72.61***
Growth	0.33***	.02	0.46***	.01	22.22***
Speed					
Intercept	-0.69***	.05	-0.86***	.04	7.53**
Growth	-0.50***	.02	-0.21***	.02	85.62***
Inhibition					
Intercept	0.91***	.07	1.35***	.05	23.98***
Growth	0.49***	.02	0.57***	.02	4.91*
Phonological loop					
Intercept	-0.02	.04	0.32***	.03	44.54***
Growth	0.08**	.01	.10***	.01	1.03
Sketchpad					
Intercept	0.15*	.07	0.53***	.05	19.04***
Growth	0.11***	.03	0.25***	.02	11.34***
Executive					
Intercept	0.36***	.05	0.81***	.04	38.78***
Growth	0.20***	.02	0.31***	.02	14.27***

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

measures of fluency and STM (phonological loop). An advantage in growth was found for the at-risk group on measures of phonological knowledge and reading. However, when the total sample was considered, growth in the phonological loop and the central executive component of WM were related to growth in problem solving accuracy. In addition, performance levels on all three components of WM were significantly related to problem solving accuracy.

### *Hierarchical Regression*

The final analysis determined those cognitive and achievement measures in Wave 1 that best predicted problem solving in Wave 3. These analyses specifically addressed the question as to which components of WM were predictive of problem solving accuracy and whether those predictions were mediated by individual differences in knowledge base, phonological processing, and/or a central executive system. Also included in this part of the analysis for the comparative purposes were the Wave 3 criterion measures of math calculation and knowledge of problem solving components. The estimates for these measures are shown in Table 7.

Prior to the regression analysis, we examined the correlations among the latent measures and the age variable. Correlations between scores for Waves 1 and 3 are shown in Table 8. The majority of

correlations greater than .20 in magnitude were significant at the .001 alpha level. As shown, all Wave 1 measures were significantly related to Wave 3 measures of problem solving accuracy, math calculation, and knowledge of the problem solving components.

An inspection of Table 8 revealed three important findings. (To interpret the results, we considered  $r_s > .40$  as substantial correlations.) First, problem solving accuracy in Wave 3 was substantially related to Wave 1 measures of reading, phonological knowledge, fluency, speed, phonological loop, and executive processing. Second, calculation skill and knowledge of problem solving components in Year 3 were strongly correlated with all Wave 1 measures except for inhibition. Finally, high correlations emerged between reading in Wave 1 and measures of math calculation and knowledge of problem solving components in Wave 3.

### *Predictions of Problem Solving*

We investigated whether the relationship between problem solving in Wave 3 and WM components in Wave 1 was maintained when blocks of variables such as knowledge base (e.g., calculation skill), phonological processing (STM, phonological knowledge), and age were entered into the regression analysis. For comparative purposes, we also considered in the analysis Wave 3 criterion

Table 6  
*Contribution of Working Memory Growth to Problem Solving*

Conditional model	Estimate	SE	t
Fixed effects			
Intercept	.47	.02	17.49***
Linear growth	.21	.01	14.51***
Working memory			
Intercept			
STM	.14	.03	3.50**
Executive	.09	.02	3.23***
Sketchpad	.10	.02	4.50***
Growth (linear)			
STM	-.05	.02	-2.08*
Executive	-.10	.01	-5.51***
Sketchpad	.005	.01	0.36
	Variance estimate	SE	z
Random effects (Participants $\times$ Teachers)			
Intercept	.07	.01	5.55**
Growth (linear)	.01	.004	3.95**
Residual	.09	.008	13.62***

Note. STM = short-term memory.

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

measures of math calculation ability and knowledge of problem solving components.

For our first set of analyses, we determined the amount of variance in problem solving performance in Wave 3 that was accounted for by WM components alone in Wave 1. For each model, we entered predictor variables from Wave 1 into the equation simultaneously, so that beta values reflected unique variance (the influence of all other variables was partialled out). For Model 1, we assumed that if scores related to STM (phonological loop) and visual-spatial sketchpad were partialled out in the regression analysis, then the residual variance related to WM could be attributed to the central executive system (e.g., Engle et al., 1999). As shown in Model 1 in Table 9, all components of WM contributed significant variance to the Wave 3 criterion measures. Performance related to the components of WM in Wave 1 contributed to approximately 36% of the variance to problem solving, 37% of the variance to arithmetic calculation, and 36% to problem solving component knowledge in Wave 3. As a follow-up to this model, we considered percentage of variance accounted for in the predictions when only the central executive component of WM was entered into the regression model. The central executive component of WM accounted for 27% of the variance for problem solving accuracy ( $R^2 = .27$ ),  $F(1, 291) = 110.06$ ,  $p < .001$ , 28% for math calculation ( $R^2 = .28$ ),  $F(1, 291) = 112.06$ ,  $p < .001$ , and 29% for problem solving knowledge ( $R^2 = .29$ ),  $F(1, 291) = 116.31$ ,  $p < .001$ . Thus, the contribution of the storage components of WM (phonological loop and visual-spatial sketchpad) in Model 1 increased the percentage accounted for by 9%, 9%, and 7%, respectively, when predicting problem solving accuracy, calculation accuracy, and problem solving knowledge.

In Model 2, we determined whether the contribution of the component scores related to WM in predicting the criterion measures were merely a function of chronological age at Wave 1. As shown in Table 9, all components of WM maintained significant

predictions of the criterion measures when age was partialled out in the analysis. The predictor variables in Model 2 for Wave 1 contributed approximately 38% of variance in word problem solving accuracy, 52% of the variance in math calculation, and 71% of the variance in problem solving component knowledge.

### *Role of Executive Processes*

As shown in Table 9, Model 3 assessed whether the contribution of WM to problem solving remained significant when cognitive

Table 7  
*Standardized Estimates Used to Create Latent Measures in the Regression Analysis*

Domain	Estimate	t
Problem Solving Accuracy (Wave 3)		
Mental Computation: WISC-III	.96	10.14***
Word Problems: Semantic Structure	.25	2.75**
Math (Wave 3)		
WRAT	.92	5.32***
WIAT	.96	4.81***
Computational Fluency	.42	15.58***
Reading (Wave 1)		
WRAT	.93	6.48***
TOWRE: real words	.96	5.94***
Comprehension: WRMT	.94	6.08***
Problem Solving Components (Wave 3)		
a. Question	.78	9.21***
b. Numbers	.83	8.25***
c. Goal	.81	9.05***
d. Operations	.89	9.88***
e. Algorithm	.90	9.06***
Phonological Knowledge: Wave 1		
Pseudowords: TOWRE	.91	9.82***
Elision (Segmentation): CTOPP	.71	7.65***
Naming Speed: Wave 1		
Digit Naming Speed: CTOPP	.95	6.79***
Letter Rapid Naming: CTOPP	.94	6.71***
Phonological Loop (STM): Wave 1		
Digit Forward: WISC-III	.37	12.48***
Pseudoword Span	.45	9.40***
Real Word Span	.72	18.55***
Digit Backward: WISC-III	.42	6.08***
Executive System (working memory): Wave 1		
Update	.58	25.80***
Listening Sentence Span	.50	23.07***
Digit/Sequence Span	.49	24.48***
Semantic Association	.30	14.86***
Visual-Spatial Sketchpad: Wave 1		
Visual Matrix	.94	45.65***
Mapping/Direction	.14	6.29***
Inhibition: Wave 1		
Random Generation Letters	.58	6.59***
Random Generation Numbers	.84	9.66***
Fluency: Wave 1		
Categorical Fluency	.57	5.03***
Letter Fluency	.61	7.15***
Age	1.00	142.22***

Note. Standardized factor loadings represent partial regression coefficients. STM = short-term memory; WISC-III = Wechsler Intelligence Scale for Children—Third Edition; WRAT = Wide Range Achievement Test; WIAT = Wechsler Individual Achievement Test; TOWRE = Test of Word Reading Efficiency; CTOPP = Comprehensive Test of Phonological Processing.

All loadings were statistically significant: \*\*  $p < .01$ . \*\*\*  $p < .001$ .



Table 8  
*Intercorrelations Among Mathematics, Reading, and Cognitive Processing Variables*

Variable	1	2	3	4	5	6	7	8	9	10	11	12
Year 1												
1. Age	—											
2. Reading	.62	—										
3. Phonological Knowledge	.48	.89	—									
4. Fluency	.32	.49	.50	—								
5. Inhibition	-.09	-.13	-.11	.01	—							
6. Speed	-.52	-.73	-.67	-.43	.16	—						
7. Sketch Pad	.28	.32	.27	-.20	-.03	-.24	—					
8. Phonological Loop	.24	.47	.45	.32	-.08	-.42	.11	—				
9. Executive	.36	.58	.60	.43	-.16	-.45	.23	.52	—			
Year 3												
10. Problem Solving Accuracy	.37	.59	.55	.40	.01	-.43	.36	.41	.53	—		
11. Calculation	.60	.80	.68	.42	-.03	-.60	.34	.45	.53	.66	—	
12. Problem Knowledge <sup>a</sup>	.78	.79	.64	.44	-.06	-.61	.33	.44	.54	.57	.76	—

<sup>a</sup> Knowledge of problem solving components.

$r_s > .21, p < .0001$ .

variables (fluency, inhibition) assumed to underlie executive processing were entered into the analysis. That is, if the central executive system of WM was primarily related to inhibition and/or fluency, then entering these variables into the regression model should have partialled out the influence of the central executive component of WM. Model 3, when compared with Model 2, increased the percentage of variance accounted for by 2% in predicting word problem solving accuracy, by 6% in predicting calculation, and by 2% in predicting problem solving component knowledge. Wave 1 variables that contributed unique variance to all three-criterion measures were components of WM related to the central executive system. Chronological age contributed unique variance to calculation and problem solving component knowledge but not problem solving accuracy. In general, we did not find support for the notion that measures of fluency, speed, and inhibition eliminated the significant contribution of the central executive system to problem solving.

#### *Role of Phonological Processes*

Model 4, as shown in Table 9, determined whether the variables related to reading and phonological processing (e.g., reading, phonological knowledge) partialled out the influence of WM in predicting problem solving accuracy, math calculation, and knowledge of problem solving components. When compared with Model 3, Model 4 increased the percentage of variance accounted for by 4% for problem solving accuracy, by 11% for calculation, and by 5% for knowledge of problem solving components. The results showed that when compared with Model 3, the influence of naming speed was partialled out with the entry of reading variables into the equation. The only Wave 1 variable that contributed unique variance to all three-criterion measures was reading. Chronological age contributed unique variance to calculation and problem solving component knowledge. As found in Model 3, the central executive component of WM contributed unique variance to word problem solving accuracy and knowledge of problem solving components. In general, the results suggest that when compared with the previous models, the significant effects of the central

executive system of WM were eliminated only on the measure of calculation when measures of reading were entered into the regression analysis.

#### *Role of Knowledge Base*

Admittedly, we were surprised that measures of the central executive component of WM were not eliminated in predicting later performance of problem solving accuracy when measures of reading and executive processing activities (e.g., inhibition) were entered into the analysis. We reasoned that, perhaps, domain-specific knowledge in math calculation and general reasoning ability may account for the robust findings related to the WM measures. Thus, in a final model (Model 5) we entered Wave 3 scores related to math calculation ability and knowledge of problem solving components, and Wave 1 scores for fluid intelligence (Raven Colored Progressive Matrices Test) into the regression model. This model directly tested whether individual differences in children's knowledge base underlie the relationship between WM and problem solving. The results are shown in Table 10. The model determined whether domain-specific knowledge (math) and reasoning ability (fluid intelligence) would partial out the effects of WM in predicting problem solving accuracy. When compared with Model 4, the final model increased the percentage of variance accounted for by 5% when predicting problem solving accuracy. Predictor variables in Wave 3 that contributed significant variance to problem solving accuracy were math calculation and problem solving knowledge, and predictor variables in Wave 1 that contributed significant variance were fluid intelligence, visual-spatial sketchpad, and the central executive component of WM. Interestingly, when compared with Model 4, Model 5 completely eliminated the significant influence of reading on problem solving accuracy.

In summary, there were two important findings related to the hierarchical regression analysis. First, entering various cognitive variables attributed to executive processing (speed, fluency, and inhibition) as well as reading skill into the regression model did not eliminate the contribution of executive processing to predicting

Table 9

*Hierarchical Analysis Predicting Word Problem Solving, Math Calculation, and Problem Solving Component Knowledge in Wave 3 From Cognitive and Achievement Variables in Wave 1*

Regression modeling	Word problem solving				Calculation				Problem solving components			
	<i>B</i>	<i>SE</i>	$\beta$	<i>t</i>	<i>B</i>	<i>SE</i>	$\beta$	<i>t</i>	<i>B</i>	<i>SE</i>	$\beta$	<i>t</i>
Model 1												
Sketchpad	.26	.04	.26	5.47***	.23	.04	.23	4.77***	.22	.04	.22	4.49***
Phonological loop	.20	.06	.17	3.02***	.37	.06	.22	3.92***	.25	.06	.20	3.55***
Executive	.48	.07	.37	6.63***	.47	.07	.36	6.39***	.49	.07	.38	6.70***
Model 2												
Age	.14	.04	.14	2.75***	.43	.04	.44	9.89***	.63	.03	.65	18.69***
Sketchpad	.23	.04	.23	4.80**	.14	.04	.13	3.22***	.08	.03	.08	2.44*
Phonological loop	.19	.06	.15	2.81**	.22	.05	.17	3.67***	.17	.04	.14	3.36*
Executive	.44	.07	.33	5.90***	.32	.06	.25	4.93**	.28	.04	.22	5.41**
Model 3												
Fluency	.14	.07	.10	1.94	.10	.06	.08	1.75	.11	.04	.08	2.45*
Speed	-.10	.06	-.11	-1.83	-.23	.05	-.23	-4.34***	-.14	.04	-.14	-3.50**
Inhibition	.08	.08	.06	1.28	-.02	.07	-.01	-0.35	-.01	.05	-.01	-0.38
Age	.08	.05	.08	1.54	.34	.04	.35	7.75***	.57	.03	.59	16.13***
Sketchpad	.21	.04	.22	4.42**	.11	.04	.11	2.71**	.05	.03	.05	1.67
Phonological loop	.16	.06	.12	2.21***	.14	.05	.15	2.35**	.11	.04	.09	2.38**
Executive	.37	.07	.29	4.78***	.23	.06	.17	3.32**	.20	.04	.16	3.84**
Model 4												
Reading	.34	.12	.34	2.75**	.68	.09	.68	7.29***	.52	.07	.53	6.85***
Phonological knowledge	-.01	.11	-.01	-0.01	-.09	.08	-.09	-1.15	-.18	.06	-.17	-2.56*
Fluency	.07	.07	.06	1.21	.02	.05	.01	0.38	.07	.04	.05	1.60
Speed	.02	.06	.02	0.36	.01	.05	.01	0.24	.002	.04	.002	0.05
Inhibition	.07	.08	.05	1.08	-.04	.06	-.03	-0.96	-.03	.05	-.02	-0.88
Age	-.01	.05	-.01	-0.17	.17	.04	.17	3.81**	.45	.03	.46	12.40***
Sketchpad	.20	.04	.20	4.16***	.08	.03	.08	2.25*	.03	.03	.03	1.13
Phonological loop	.13	.06	.10	1.84	.09	.05	.07	1.68	.09	.04	.06	1.75
Executive	.27	.08	.21	3.40**	.06	.05	.04	1.00	.11	.05	.09	2.30*

Note: Model 1,  $F(3, 289) = 54.82, p < .001, R^2 = .36$ ;  $F(3, 289) = 55.51, p < .001, R^2 = .37$ ;  $F(3, 289) = 54.20, p < .001, R^2 = .36$ . Model 2,  $F(4, 288) = 43.94, p < .001, R^2 = .38$ ;  $F(4, 288) = 80.05, p < .001, R^2 = .52$ ;  $F(4, 288) = 176.92, p < .001, R^2 = .71$ . Model 3,  $F(7, 284) = 27.30, p < .001, R^2 = .40$ ;  $F(7, 284) = 52.62, p < .001, R^2 = .57$ ;  $F(7, 284) = 110.75, p < .001, R^2 = .73$ . Model 4,  $F(9, 282) = 24.23, p < .001, R^2 = .44$ ;  $F(9, 282) = 65.72, p < .001, R^2 = .68$ ;  $F(9, 282) = 111.38, p < .001, R^2 = .78$ .

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

problem solving accuracy or performance on the knowledge of problem solving components task. Second, performance on different components of WM predicted different criterion measures. The executive system and visual-spatial sketchpad were important predictors of problem solving accuracy, whereas performance related to the visual-spatial sketchpad predicted calculation, and the executive system predicted performance on the knowledge of problem solving components task.

### Discussion

The purpose of this study was to identify the cognitive processes in WM that predict mathematical problem solving performance in elementary school children who are at risk and not at risk for serious math difficulties. We determined whether WM was a valid construct in the predictions of problem solving and whether measures of WM predicted problem solving above and beyond children's performance on measures related to reading, phonological processing, computational knowledge, and domain-specific knowledge. We also determined whether growth in WM was related to growth in problem solving accuracy. Overall, our data provide substantial support for the notion that the capacity to store and process material in WM significantly constrains a child's

ability to problem solve during the elementary school years. Further, children identified at risk for problem solving difficulties in Wave 1 maintained their risk status on measures of problem solving and WM 2 years later. In the discussion that follows we review our findings related to WM and problem solving. We then discuss the implications of our findings for future research and practice.

### Testing of Three Models

We tested three models of WM and its influence on growth in problem solving. One model tests whether phonological processes (e.g., STM-phonological loop, phonological knowledge) play a major role in predicting performance in problem solving and whether the phonological system mediates the influence of executive processing (WM) on problem solving. Phonological processes in this study were related to latent measures of STM and phonological knowledge (measures of elision, pseudoword reading). The model follows logically from the reading literature that links phonological skills to new word learning (e.g., Baddeley et al., 1998), comprehension (Perfetti, 1985), and mental calculation (e.g., Logie, Gilhooly, & Wynn, 1994). The model assumes that poor problem solvers have deficits in the processing and storage of



Table 10  
*Predictions of Year 3 Problem Solving Accuracy Based on Wave 3 Math Calculation, Problem Solving Knowledge, and Wave 1 Fluid Intelligence, Reading, and Cognitive Variables*

Model 5 predictor	B	SE	$\beta$	t
Wave 3				
Problem Solving Knowledge	.25	.11	.12	2.13*
Calculation	.30	.08	.27	3.42**
Wave 1				
Fluid Intelligence (Raven)	.13	.04	.16	2.85**
Reading	.12	.12	.12	1.00
Phonological Knowledge	-.01	.10	.10	-0.09
Fluency	.02	.07	.007	0.33
Speed	-.004	.06	-.004	-0.06
Inhibition	.09	.06	.07	1.60
Age	-.15	.06	-.16	-2.39*
Sketchpad	.15	.04	.14	3.23***
Phonological Loop	.12	.06	.09	1.85
Executive	.19	.08	.15	2.34*

Note.  $F(12, 279) = 22.52, p < .001, R^2 = .49$ .

\*  $p < .05$ . \*\*  $p < .01$ .

phonological information that creates a bottleneck in the flow of information to higher levels of processing.

The second model suggests that domain-specific knowledge in LTM mediates individual differences in WM and problem solving. Measures of LTM in this study were related to calculation ability, reading, and knowledge of problem solving components. It has recently been argued that some of the functions of the central executive system are to access information from LTM (e.g., Baddeley & Logie, 1999), and, therefore, this model suggests that controlling for the activation of LTM (e.g., arithmetic calculation, knowledge of algorithms) would partial out the influence of WM on problem solving.

The final model suggests that problem solving performance relates to executive processing, independent of the influence of the phonological system and LTM. This assumption follows logically from the problem solving literature suggesting that abstract thinking, such as comprehension and reasoning, requires the coordination of several basic processes (e.g., Engle et al., 1999; Just, Carpenter, & Keller, 1996; Kyllonen & Christal, 1990). We assumed that measures of executive processing in this study were related to latent measures of WM, and measures assumed to reflect some of the activities of the central executive system were inhibition (random generation of letters and numbers) and activation of LTM (composite measures of reading, arithmetic calculation, knowledge of problem solving components). We also assumed based on the work of Engle et al. (1999) and others (e.g., Cowan, 1995) that after storage processes (phonological loop and visual-spatial) were partialled from the analysis that the residual variance related to WM captured a key process of the central executive system referred to as *controlled attention*. The results yield two clear findings in support of the third model. However, when interpreting these findings the reader needs to keep in mind that other processes not reflected in our latent measures may also play a role.

First, WM contributes unique variance to problem solving beyond what phonological processes (e.g., phonological knowledge), reading skill, inhibition, and processing speed contribute. The

results show that WM performance in Wave 1 contributed approximately 36% of the variance to problem solving accuracy in Wave 3 when entered by itself in the regression analysis. Thus, there is clear evidence that multiple systems of WM contribute important variance to problem solving performance beyond processes related to speed, phonological knowledge, and reading skill.

Second, WM performance predicted problem solving accuracy when the hierarchical regression analysis included measures of LTM. Measures of LTM in this study comprised the tasks related to reading, calculation, and knowledge of problem solving components. It has been argued that some of the functions of the central executive system include accessing information from LTM (e.g., Baddeley & Logie, 1999; Conway & Engle, 1994). Our results suggest, however, that although WM tasks draw information from LTM (e.g., knowledge of components related to problem solving), it may be the controlled attention (monitoring of attention) component of WM that plays a more important role in mathematical problem solving growth. The results clearly show in the final model (see Table 10, Model 5) that neither knowledge of problem solving components nor calculation ability partialled out the significant influence of WM in predicting mathematical problem solving accuracy. We will now address the three specific questions that directed our study.

1. Do children identified at risk or not at risk in Wave 1 vary on measures of growth in problem solving and WM across the testing waves?

To answer this question, we estimated intercept values in Wave 3 and the rate of growth across testing waves on measures of problem solving and WM. The growth curve analysis showed that ability group differences in problem solving emerged for the estimated intercept values at Wave 3. Thus, although ability group differences would be expected at Wave 1, the results showed that these differences were sustained at Wave 3. Although we were uncertain whether improvement in reading would mitigate any word problem solving differences in subsequent years, children at risk for SMD in Wave 1 were also at risk in Wave 3. In addition, significant differences emerged between the two groups in linear growth, with the at-risk group showing a lower growth rate when compared with the not-at-risk group. We also found that the intercept and linear growth parameters were significantly different between children at risk for SMD and not at risk on measures of math calculation and the executive component of WM. Children at risk for SMD had lower intercepts and smaller growth rates than children not at risk. Interestingly, no differences were found between the ability groups on growth estimates for performance on measures of the phonological loop and fluency.

We also compared the two groups across a broad array of achievement and cognitive measures. The general pattern in the repeated measures ANCOVA for the intact sample across all three waves was that the comparisons between at-risk and not-at-risk students were significant at each wave and for each academic and cognitive domain. However, greater differences emerged between ability groups for some domains than others. A comparison of groups across achievement and cognitive domains (collapsed across time) indicated that greater group differences in favor of children not at risk emerged on measures of problem solving accuracy, math calculation, reading, and speed when compared

with the cognitive variables (fluency, inhibition, and memory components; e.g., phonological loop) and knowledge of problem solving components. The general trend was that children not at risk scored higher than those at risk for SMD, and higher difference scores emerged for academic domains (e.g., reading, math calculation) than cognitive domains.

2. Is growth in WM related to growth in word problem solving accuracy?

Our multilevel analysis showed a positive effect for growth in WM and problem solving accuracy. The general findings of the growth analysis were that ability group effects related to intercepts and growth parameters emerged for measures of problem solving. Children at risk for SMD showed lower levels of performance and less linear growth rate than children not at risk. When the total sample was analyzed, however, another important finding was that all intercept values related to components of WM (phonological loop, visual-spatial sketchpad, and central executive) were significantly related to problem solving. A significant relationship was found for the level of performance on all three WM components to problem solving accuracy. However, only the linear growth in the phonological loop and executive processing was related to word problem solving.

No doubt, it could be argued that growth rates in problem solving are merely a function of classroom instruction. We addressed this issue by comparing an unconditional and conditional model that included growth of the students nested within classroom with and without the inclusion of WM. The conditional model that included WM reduced the influence of classroom (teacher) variance on between-children intercepts by 61% and between-children growth by 66%. Thus, we found that a large percentage of the variance related to participant performance nested within classroom instruction on problem solving could be accounted for by individual differences in WM.

3. Which components of WM were predictive of problem solving accuracy and are those predictions mediated by individual differences in knowledge base, phonological processing, and/or the central executive system?

We tested this question by utilizing a hierarchical regression analysis, in which variables related to WM, cognitive processing (speed, inhibition, fluency), and achievement (e.g., reading, phonological knowledge) in Wave 1 were systematically entered into the regression equation to predict problem solving in Wave 3. There were two important findings.

First, the results show that the executive component of WM contributes unique variance to problem solving accuracy beyond what phonological processes (e.g., STM, phonological knowledge) contribute. The results show that the executive component of WM performance in Year 1 contributes approximately 27% of the variance to problem solving accuracy in Wave 3 when entered by itself in the regression analysis.

Second, the executive component of WM performance contributes unique variance to problem solving accuracy even when the hierarchical regression includes measures of LTM. Our results suggest, however, that although WM tasks draw information from LTM (e.g., knowledge of components related to problem solving),

it may be the monitoring of the components in WM system (e.g., manipulation of information) and storage of visual-spatial information that play a more important role in predicting later performance on mathematical problem solving measures. However, our results must be clarified because we have merely partialled out the influence of the amount of knowledge available. A distinction must be drawn between the amount of WM resources available to produce activation and the amount of knowledge that can be activated. Thus, in our regression analysis we assessed the quality of information available, not the capacity to activate LTM knowledge. The results merely show that both knowledge of problem solving components and calculation ability in Wave 3 did not partial out the significant influence of WM in Wave 1 in predicting word problem solving accuracy in Wave 3.

As a parallel to predicting word problem solving accuracy, we also considered predictions related to mathematical computation. In terms of math calculation in Wave 3, the best Wave 1 predictors were reading, chronological age, and the visual-spatial storage (sketchpad) component of WM. Thus, in contrast to the important role that the executive component of WM plays in problem solving accuracy, it appears that performance in the visual-spatial sketchpad also predicts math calculation. The visual-spatial sketchpad is specialized for the processing and storage of visual material, spatial material, or both and for linguistic information that can be recoded into visual forms (see Baddeley, 1986, for a review). Our findings are consistent with studies (Gathercole & Pickering, 2000a, 2000b) finding that visual-spatial WM abilities, as well as measures of executive processing, were associated with attainment levels on a national curriculum for children ages 6 to 7 years. Children who showed marked deficits in curriculum attainment also showed marked deficits in visual-spatial WM. Thus, there is a strong relationship between measures of the visual-spatial sketchpad and academic performance in the younger grades.

### *Comparisons With Related Studies*

Not all studies have found a significant relationship between WM and problem solving. For example, Fuchs et al. (2006) studied the cognitive correlates of arithmetic computation and arithmetic word problem ability of third graders. Predictor variables were language, nonverbal problem solving, concept formation, processing speed, long-term memory, WM, phonological decoding, and sight word reading efficiency. Word problem accuracy was best predicted by concept formation and sight word efficiency. The only significant cognitive predictors of arithmetic were phonological decoding and processing speed. Likewise, earlier work by Fuchs et al. (2005) showed that phonological processing was a unique determinant of the development of arithmetic skills in first graders. The study by Hecht, Close, and Santisi (2003) also demonstrated that phonological processing almost completely accounted for the association between reading and computational skill in older children. In contrast, Lee, Swee-Fong, Ee-Lyn, and Zee-Ying (2004) assessed the performance of 10-year-olds on measures of word problems, WM, intelligence, and reading ability. They found that children who had greater storage and greater central executive system capacities related to WM measures were better able to solve mathematical problems. At the same time they found that children with higher IQs, reading proficiencies, and vocabularies, performed better on mathematical problems. Further,



although WM was found to be related to problem solving, WM did not contribute important variance to problem solving accuracy after measures of reading were entered into the regression analysis. This finding replicated an earlier finding from the study by Swanson et al. (1993), who found that reading comprehension performance superseded any unique contributions made by WM measures to problem solving (also see Kail & Hall, 1999, for a similar finding). These findings are in contrast to other researchers (e.g., De Beni, Palladino, Pazzaglia, & Cornoldi, 1998; Passolunghi, Cornoldi, & De Liberto, 2001), who have shown that when controlling for vocabulary and reading, poor problem solvers exhibit lower WM scores (as indexed by their poorer inhibitory abilities) than do good problem solvers.

Thus, there is some confusion in the literature about the role of WM in problem solving. To examine these issues, we give attention to the recent study by Fuchs et al. (2006). As stated, Fuchs et al. (2006) failed to find that WM predicted word problems accuracy, and this finding varies considerably from our findings. There are three major distinctions between the Fuchs et al. (2006) study and our study. First, there were variations in WM measures. Fuchs et al. (2006) primarily relied on two WM measures. One was the backward span that some consider a STM rather than WM measure (Colom, Abad, et al., 2005; Colom, Flores-Mendoza, et al., 2005; also see Footnote 1). In contrast, our study used a comprehensive battery of WM measures. Second, participants in the Fuchs et al. (2006) study were primarily sampled from Title 1 schools and identified on the basis of the performance on a nonstandardized test of computational fluency. Further, approximately 60% of their sample received subsidized lunch. In contrast, the present study sample of children comes from mostly middle-class or upper middle-class homes who were selected on a standardized measure of problem solving. In addition, it is important to note that the at-risk sample in the present study included students who were within the average range in arithmetic and reading skills. In contrast to the Fuchs et al. (2006) study that defined risk as the ability to calculate, our study defined risk as the ability to problem solve and name numbers quickly. Finally, as indicated by one reviewer, our results when compared with the work of Fuchs et al. (2006) were primarily due to methodology differences. For example, variations emerged in the criterion measure. In the Fuchs et al. (2006) study, word problems were presented as text and students had to process text. Children were allowed 3 s to respond after hearing the problem but could reread the passages if necessary. This activity would certainly lessen the demands on WM capacity because children can reread the sentence. In contrast, we read the problem to the students and students were able to focus on the problem being read while maintaining information in WM. Therefore, the reviewer was not surprised that reading played a minimal role in our study and was important in the Fuchs et al. (2006) study. Further, as indicated by the reviewer, the presentation of word problems in text in the Fuchs et al. (2006) study may be more ecologically valid than ours. We would argue that it is equally valid to problem solve without recourse to text—thinking on your feet as it were. Everyday problems are not always presented in text and it may be ecologically valid to also teach problem solving without having to rely on printed text. As indicated by an anonymous reviewer of this article, we could argue that we have a purer

measure of word problem solving because reading was not required.

On the issues of reading, some studies have argued that reading proficiency mediates the relationship between WM and problem solving. For example, Lee et al. (2004) showed that literacy measures provided greater contribution than measures related to the central executive system to word problem solving ability in children. Although WM was significantly related to word problem solving ability, its variance was significantly reduced when the influence of reading was partialled out. Further, they found that the storage component of WM, primarily the phonological loop and the visual-spatial sketchpad, did not contribute directly to mathematical performance.

Similar to Lee et al. (2004), as well as Fuchs et al.'s (2006) study, we found that reading was an important predictor of calculation. However, the contribution of WM to problem solving in our study was larger than that of reading. No doubt, the weak association of reading to problem solving in Model 5 (see Table 9) could be explained by the fact that all problems used in the current study were read to the students; consequently, reading was weakened as a possible mediating variable. We argue, however, that if children are unable to understand and decode a question, further processing is unlikely to lead to a correct solution. Therefore, reading abilities may account for the similarities and differences in the findings across studies.

In addition, there are other studies that clearly show that reading or reading-related processes do not directly mediate the influence of WM on problem solving. For example, Swanson and Sachse-Lee (2001) found with 12-year-old math-disabled and chronologically age-matched peers that phonological processing, WM (executive component), and visual WM contributed unique variance to problem solving accuracy. Thus, they did not find support that reading ability or literacy processes mediated the role of WM in the problem solving accuracy. In a follow-up study, Swanson (2004) compared two age groups (7 and 11) on WM and problem solving measures. This study found that regardless of age, WM predicted problem solving accuracy in word problems independent of measures of problem representation, knowledge of operations and algorithms, phonological processing, fluid intelligence, and achievement in reading and math. Further, the results suggest that a central executive system underlies age-related improvements in word problem solving accuracy.

In general, we argue that the correct calculation of numbers presented on paper requires some minimum threshold of reading ability. A lack of control for this confound may lead to attributing difficulties in math to the same cognitive processes (phonological process) as reading. In addition, children who read at a certain threshold (e.g., above the 25th percentile in word identification) may fail to be adequately diagnosed as having potential difficulties in mathematical problem solving. In order to circumvent these problems, we feel it is necessary to identify processes in children at risk for SMD that are distinct from the processes related to reading. Our results clearly show that children at risk for SMD on a nonreading measure but whose performance is in the average range on measures of fluid intelligence, reading skill, and math calculation skill are deficient on measures of WM when compared with children not at risk for problem solving.



### Implications

In general, our findings show that WM performance predicts later performance in problem solving accuracy, mathematical calculation, and knowledge of problem solving components. What are the implications of our findings to education of children? We believe the results have three implications.

1. Our classification criteria have predictive validity. Children identified at risk for SMD in Wave 1 were deficient in problem solving in Wave 3. Children identified at risk for SMD in Wave 1 did not catch up to children not at risk in Wave 3 in problem solving, math, and WM. Further, as shown in Figure 1, several other areas were found deficient in children identified at risk for serious mathematical problem solving difficulties (e.g., reading, random generation). Thus, our classification based on digit naming speed and the Arithmetic subtest of the WISC-III (which includes word problems) was a valid discriminator in predicting subsequent growth in Year 3.

Of course, one may criticize our classification criteria both in terms of the tasks we used, especially the Arithmetic subset of the WISC-III, and the cutoff criteria as well. Let us consider three arguments related to the Arithmetic subtest. One argument is that risk status should be based on calculation skill and not problem solving ability. However, we argue that we were able to identify a substantial number of children who were in the average range on reading and calculation measures but had clear difficulties listening to a problem with a verbal text and coming up with a correct solution. We are very much aware of the work by others (e.g., Geary, 2003), indicating that math disability should focus on calculation only. However, selection criterion for risk has varied across studies from the 25th percentile to the 45th percentile (see Swanson & Jerman, 2006, for a review). We think one of the reasons for this variation in cutoff scores is that standardized calculation scores have an extremely restrictive range of items for the younger grades and/or items do not match demands of the classroom. That is, popular standardized measures of calculation place children higher than actual classroom performance (e.g., see Fuchs et al., 2004, p. 496, for discussion of this issue). However, when problem solving difficulties were considered, we were able to sample enough children below a conservative cutoff score of the 25th percentile.

Another argument is that the Arithmetic subtest from the WISC-III was not intended as a measure of math. Although Sattler (2001) and others suggested that the Arithmetic subtest from the WISC-III was not intended to be a math measure, it is clearly correlated with math achievement, as well as vocabulary and other related skills. Regardless of the intent of the measure, the arithmetic subtest requires children to answer simple to complex problems involving arithmetic concepts and numerical reasoning. The first couple of problems require the direct counting of discrete objects. Problems 3, 4, and 5 require subtraction using objects and stimuli. The remaining problems require addition, subtraction, multiplication, and division.

Another argument is that the Arithmetic subtest itself is a WM measure. No doubt this task taps WM as well as prior learning, but it also correlates more highly with information and similarities than the other subtests (e.g., Digit Span). In fact, it has a *g* loading the same as Vocabulary, Information, and Similarities (see Keith et al., 2006, p. 122), whereas Digit Span is substantially lower.

Although we recognize that the mental computation required for this task does place constraints on WM capacity, the goal in our study was to find out which components of WM are the most constrained for children who have difficulties related to mental calculation.

In summary, our finding is important because children who were at risk for problem solving were generally in the average range on reading and math calculation measures but sustained difficulties on a number of measures across all testing waves.

2. A major cognitive component that underlies risk status in elementary school children with average intelligence is the executive component of WM. In contrast to the work on reading, it does not appear to be the case in this study that deficits in WM were merely a manifestation of deficits in the phonological system (also see Swanson & Ashbaker, 2000, for a similar finding). We found in our regression modeling that WM predicted math problem solving even when the influence of STM was partialled out.

The above finding is consistent with other studies showing that WM is an important predictor of various cognitive operations even when the influence of STM is partialled out in the analysis (e.g., Engle et al., 1999). Although we assumed that the residual variance related to WM was related to some aspect of the central executive system, most notably controlled attention, an important point raised by one anonymous reviewer of this article is that we have no direct measure indicating that the residual variance related to WM is controlled attention. Further, the reviewer also indicated that our findings are merely an artifact of *g*. For example, children at risk for SMD had lower Raven scores than those not at risk. However, it is important to note that fluid intelligence was partialled out in the analysis in Model 5. Even under these conditions, WM still contributed significant variance to word problem solving. However, the reviewer is correct that one cannot assume that the residual variance related to WM is controlled attention. Let us consider this point in more detail.

Our results extend Engle et al.'s (1999) findings to children that when controlling for the correlations between WM and STM, the residual variance for the WM factor predicts problem solving. The question emerges, however, as to whether the residual variance attributed to WM and problem solving reflects controlled attention (e.g., Engle et al., 1999), a domain general attentional resource involved in the activation of information from LTM (e.g., Cantor & Engle, 1993), a general monitoring system that coordinates the flow of information but draws from specialized storage systems (e.g., Baddeley & Logie, 1999), or a limited-capacity resource that supports both processing and storage in a domain-specific system (e.g., Just & Carpenter, 1992).

To answer this question, let us first consider Model 3 in Table 9. We assumed that some of the activities related to executive processing component of the WM measure (i.e., controlled attention) were related to the latent measures of fluency, random generation, and speed of processing. However, we found that entry of these measures in the regression models did not eliminate the significant contribution of WM to problem solving accuracy or calculation. What these findings suggest to us is that attentional processes dissimilar from those captured by measures of fluency, speed, and random generation mediate the residual variance related to WM in predicting problem solving. Further, because our measures were not pure measures of inhibition, we may not have adequately tested the inhibition model. Interestingly, in Model 5



(see Table 9), we found that both the executive and visual-spatial component of WM significantly predicted math problem solving when measures of literacy (reading and phonological knowledge) were entered into the regression model. Thus, weak support was found for the phonological model, because the entry of those variables into the regression analysis failed to partial out the influence of WM in predictions of problem solving. Thus, verbal forms of processing efficiency (naming speed) and content (phonological knowledge) did not adequately account for the influence of WM. On the basis of these results, we argue that if WM is made up of controlled attention and storage, then when the influence of storage is partialled out in the analysis, what is left in terms of residual variance is some form of controlled attention. However, given that the random generation, fluency tasks, and measures of LTM did not eliminate the contribution of this residual variance to word problem solving accuracy, we tentatively concluded that some yet-to-be-specified aspects of controlled attentional processing play an important predictive role in elementary school children's problem solving.

3. Limitations in WM capacity can be compensated for by improved fluency and proficiency on academic tasks (i.e., knowledge of problem solving components, reading comprehension, and math computation). Although the influence of individual differences in WM on problem solving performance is robust, this does not mean that its influence cannot be compensated for. Increased performance on measures related to math calculation and knowledge of algorithms reduced the influence of individual differences in WM on problem solving. What remains to be examined is the influence of instruction and age-related development on these processes. From a practical standpoint, the finding that WM tests do mediate problem solving outcomes should be considered when we look at high stakes examinations. Although our research suggests that literacy (reading) is important in terms of augmenting (predicting) word problem solving ability (see Model 4, Table 9), this is not the only variable that needs to be considered when compensating for demands on WM. The demands on reading are not greater than the demands on WM. In fact, the actual percentage of variance contributed to word problem solving ability is higher for the WM measures than it is for the literacy measures. The results do suggest that mathematical calculation skills are important in the study; so clearly there must be a threshold that would be extremely important. It appears that the children in this study had mastered or at least performed above the threshold and, therefore, were able to rely on other resources.

### Conclusion

In summary, growth in WM predicts growth on problem solving measures. We believe these results extend those studies on individual differences that suggest that a WM system plays a critical role in integrating information during problem solving. These models explicitly posit a dual role of WM: (a) It holds recently processed information to make connections to the latest input, and (b) it maintains information for the construction of an overall solution to problems. In terms of individual differences, children who have a large WM capacity for language can carry out the execution of various fundamental problem solving processes (such as problem representation, problem execution, etc.) with less demands on a limited resource pool than children with a smaller WM

capacity. As a result, children with a larger WM capacity would have more resources available for storage while representing the problem. On the other hand, children with a smaller WM capacity might have fewer resources available for the maintenance of information during problem solving. Further, this relationship holds (at least for children) even when the influence of phonological loop (naming speed and STM) is partialled from the analysis. Yet, WM is not the exclusive contributor to variance in problem solving ability. This study also supports previous research about the importance of reading skill in solution accuracy (see Model 4). Moreover, our findings are consistent with models of high order processing, which suggest that WM resources activate relevant knowledge from LTM but also suggest that a subsystem that controls and regulates the cognitive system plays a major role. What is unclear from previous studies is whether growth in the WM system is related to growth in problem solving. This study has provided clear evidence suggesting that growth in WM is related to growth in problem solving. Thus, we think one of the core problems children face in solving mathematical word problems relates to growth in operations ascribed to WM.

### References

- Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, 131, 567-589.
- Anderson, E. R. (1993). Analyzing change in short-term longitudinal research using cohort-sequential designs. *Journal of Consulting and Clinical Psychology*, 61, 929-940.
- Anderson, J. R., Reder, L. M., & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive Psychology*, 30, 221-256.
- Baddeley, A. D. (1986). *Working memory*. London: Oxford University Press.
- Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, 4, 417-422.
- Baddeley, A. D., Gathercole, S. E., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, 105, 158-173.
- Baddeley, A. D., & Logie, R. H. (1999). The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28-61). Cambridge, England: Cambridge University Press.
- Barrouillet, P., & Lépine, R. (2005). Working memory and children's use of retrieval to solve addition problems. *Journal of Experimental Child Psychology*, 91(3), 183-204.
- Bentler, P. M., (2005). *EQS 6 Structural equations program manual* [Computer software manual]. Encino, CA: Multivariate Software.
- Bentler, P. M., & Wu, E. J. C. (1995). *Structural equations program manual* [Computer software manual]. Encino, CA: Multivariate Software.
- Bryk, A., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.
- Bull, R., Johnston, R. S., & Roy, J. A. (1999). Exploring the roles of the visual-spatial sketchpad and central executive in children's arithmetical skills: Views from cognition and developmental neuropsychology. *Developmental Neuropsychology*, 15, 421-442.
- Byrne, B. M. (2006). *Structural equation modeling with EQS*. Mahwah, NJ: Erlbaum.
- Cantor, J., & Engle, R. W. (1993). Working-memory capacity is long-term memory activation: An individual-differences approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 1101-1114.



- Colom, R., Abad, F., Rebollo, I., & Shih, P. C. (2005). Memory span and general intelligence: A latent-variable approach. *Intelligence*, 32, 623–642.
- Colom, R., Flores-Mendoza, C., Quiroga, M. Á., & Privado, J. (2005). Working memory and general intelligence: The role of short-term storage. *Personality & Individual Differences*, 39, 1005–1014.
- Conway, A. R. A., Cowan, N., Bunting, M. F., Theriault, D., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, 30, 163–183.
- Conway, A. R. A., & Engle, R. W. (1994). Working memory and retrieval: A resource-dependent inhibition model. *Journal of Experimental Psychology: General*, 123, 354–373.
- Cowan, N. (1995). *Attention and memory: An integrated framework*. Oxford, England: Oxford University Press.
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19, 450–466.
- De Beni, R., Palladino, P., Pazzaglia, F., & Cornoldi, C. (1998). Increases in intrusion errors and working memory deficit of poor comprehenders. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 51(A), 305–320.
- Dempster, F. (1985). Short-term memory development in childhood and adolescence. In C. Brainerd & M. Pressley (Eds.), *Basic processes in memory* (pp. 209–248). New York: Springer-Verlag.
- Engle, R. W., Cantor, J., & Carullo, J. J. (1992). Individual differences in working memory and comprehension: A test of four hypotheses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18, 972–992.
- Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, 128, 309–331.
- Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102, 211–245.
- Fayol, M., Abdi, H., & Gombert, J. (1987). Arithmetic problem formulation and working memory load. *Cognition and Instruction*, 4, 187–202.
- Fletcher, J. M., Epsy, K. A., Francis, P. J., Davidson, K. C., Rourke, B. P., & Shaywitz, S. E. (1989). Comparison of cutoff and regression-based definitions of reading disabilities. *Journal of Learning Disabilities*, 22, 334–338.
- Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology*, 97, 493–513.
- Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., et al. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology*, 98, 29–43.
- Fuchs, L. S., Fuchs, D., Eaton, S., Hamlett, C. L., & Karns, K. (2000). Supplementing teacher judgments of mathematics test accommodations, with objective data sources. *School Psychology Review*, 20, 65–85.
- Fuchs, L. S., Fuchs, D., Prentice, K., Hamlett, C. L., Finelli, R., & Courey, S. J. (2004). Enhancing mathematical problem solving among third-grade students with schema-based instruction. *Journal of Educational Psychology*, 96, 635–647.
- Furst, A., & Hitch, G. (2000). Separate roles for executive and phonological components of working memory in mental arithmetic. *Memory & Cognition*, 28, 774–782.
- Gathercole, S. E. (1998). The development of memory. *Journal of Child Psychology and Psychiatry*, 39, 3–27.
- Gathercole, S. E., & Pickering, S. J. (2000a). Assessment of working memory in six- and seven-year-old children. *Journal of Educational Psychology*, 92, 377–390.
- Gathercole, S. E., & Pickering, S. J. (2000b). Working memory deficits in children with low achievements in the national curriculum at 7 years of age. *British Journal of Educational Psychology*, 70, 177–194.
- Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, 40, 177–190.
- Geary, D. C. (2003). Math disabilities. In H. L. Swanson, K. Harris, & S. Graham (Eds.), *Handbook of learning disabilities*. (pp. 199–212). New York: Guilford Press.
- Geary, D. C., Brown, S. C., & Samaranayake, V. A. (1991). Cognitive addition: A short longitudinal study of strategy choice and speed-of-processing differences in normal and mathematically disabled children. *Developmental Psychology*, 27, 787–797.
- Geary, D. C., Hoard, M. K., Byrd-Craven, J., & DeSoto, M. C. (2004). Strategy choices in simple and complex addition: Contributions of working memory and counting knowledge for children with mathematical disability. *Journal of Experimental Child Psychology*, 88, 121–151.
- Harrison, J. E., Buxton, P., Husain, M., & Wise, R. (2000). Short test of semantic and phonological fluency: Normal performance, validity, and test-retest reliability. *British Journal of Clinical Psychology*, 39, 181–191.
- Hecht, S. A., Close, L., & Santisi, M. (2003). Sources of individual differences in fraction skills. *Journal of Experimental Child Psychology*, 86, 277–302.
- Hegarty, M., Mayer, R., & Monk, C. (1995). Comprehension of arithmetic word problems: A comparison of successful and unsuccessful problem solvers. *Journal of Educational Psychology*, 87, 18–32.
- Heitz, R. P., Unsworth, N., & Engle, R. W. (2005). Working memory capacity, attention control, and fluid intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 61–78). Thousand Oaks, CA: Sage.
- Henry, L. A., & Millar, S. (1993). Why does memory span improve with age? A review of the evidence for two current hypotheses. *European Journal of Cognitive Psychology*, 5, 241–287.
- Hitch, G. J., & McAuley, E. (1991). Working memory in children with specific arithmetical learning difficulties. *British Journal of Psychology*, 82, 375–386.
- Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI: Analysis of linear structural relationships by the method of maximum likelihood* [Computer software manual]. Mooresville, IN: Scientific Software.
- Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99, 122–149.
- Just, M. A., Carpenter, P. A., & Keller, T. A. (1996). The capacity theory of comprehension: New frontiers of evidence and arguments. *Psychological Review*, 103, 773–780.
- Kail, R., & Hall, L. K. (1999). Sources of developmental change in children's word-problem performance. *Journal of Educational Psychology*, 91, 660–668.
- Keith, T., Fine, J. G., Taub, G., Reynolds, M., & Kranzler, J. (2006). High order, multisampling, confirmatory factor analysis of the Wechsler Intelligence Scale For Children—Fourth Edition: What does it measure? *School Psychology Review*, 35, 108–127.
- Keith, T., & Wittam, E. (1997). Hierarchical and cross-age confirmatory factor analysis of the WISC-III: What does it measure? *School Psychology Review*, 12, 89–107.
- Kintsch, W., & Greeno, J. G. (1985). Understanding and solving arithmetic word problems. *Psychological Review*, 92, 109–129.
- Klauer, K. J., Willmes, K., & Phye, G. (2002). Inductive reasoning: Does it transfer to fluid intelligence? *Contemporary Educational Psychology*, 27, 1–25.
- Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity? *Intelligence*, 14, 389–433.
- Lee, K., Swee-Fong, N., Ee-Lynn, N., & Zee-Ying, L. (2004). Working



- memory and literacy as predictors of performance on algebraic word problems. *Journal of Experimental Child Psychology*, 89, 140–158.
- Logie, R. H., Gilhooly, K. J., & Wynn, M. (1994). Counting on working memory in arithmetic problem solving. *Memory & Cognition*, 22, 395–410.
- Mayer, R. E., & Hegarty, M. (1996). The process of understanding mathematical problem solving. In R. J. Sternberg & T. Ben-Zeev (Eds.), *The nature of mathematical thinking* (pp. 29–54). Mahwah, NJ: Erlbaum.
- McGraw, B., & Jöreskog, K. G. (1971). Factorial invariance of ability measures in groups differing in intelligence and social economic status. *British Journal of Mathematical and Statistical Psychology*, 24, 154–168.
- Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods*, 5, 23–43.
- Miyake, A. (2001). Individual differences in working memory: Introduction to the special section. *Journal of Experimental Psychology*, 130, 163–168.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., & Howerter, A. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, 41, 49–100.
- Morris, N., & Jones, D. M. (1990). Memory updating in working memory: The role of central executive. *British Journal of Psychology*, 91, 111–121.
- National Science Foundation. (2000). *National Survey of Science and Mathematics Education*. Reston, VA: Author.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 411–421.
- Oberauer, K., Suss, H., Wilhelm, O., Wittman, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence*, 31, 167–193.
- Ornstein, P., Naus, M. J., & Liberty, C. (1975). Rehearsal and organization processes in children's memory. *Child Development*, 46, 818–830.
- Passolunghi, M. C., Cornoldi, C., & De Liberto, S. (2001). Working memory and intrusions of irrelevant information in a group of specific poor problem solvers. *Memory & Cognition*, 27, 779–790.
- Perfetti, C. (1985). *Reading ability*. New York: Oxford University Press.
- Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74, 525–556.
- Plewis, I. (1996). Statistical methods for understanding cognitive growth: A review, a synthesis and an application. *British Journal of Mathematical and Statistical Psychology*, 49, 25–42.
- Psychological Corporation. (1991). *Wechsler Intelligence Scale for Children—Third Edition*. San Antonio, TX: Harcourt Brace Jovanovich.
- Psychological Corporation. (1992). *Wechsler Individual Achievement Test*. San Antonio, TX: Harcourt Brace.
- Rasmussen, C., & Bisanz, J. (2005). Representation and working memory in early arithmetic. *Journal of Experimental Child Psychology*, 91, 137–157.
- Raven, J. C. (1976). *Colored progressive matrices*. London: H. K. Lewis.
- Reise, S. P., Widaman, K., & Pugh, R. (1993). Confirmatory factor analysis and item response theory: Two approaches for explaining measurement invariance. *Psychological Bulletin*, 114, 552–566.
- Rende, B., Ramsberger, G., & Miyake, A. (2002). Commonalities and differences in working memory components underlying letter and category fluency tasks: A dual-task investigation. *Neuropsychology*, 16, 309–321.
- Riley, M., Greeno, J. G., & Heller, J. (1983). The development of children's problem solving ability in arithmetic. In H. Ginsburg (Ed.), *The development of mathematical thinking* (pp. 153–196). New York: Academic Press.
- Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology*, 93, 346–362.
- Rosen, V. M., & Engle, R. W. (1997). The role of working memory capacity in retrieval. *Journal of Experimental Psychology: General*, 126, 211–227.
- SAS Institute. (2003). *SAS/STAT user's guide* [Computer software manual]. Cary, NC: Author.
- Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th ed.). San Diego, CA: Author.
- Shankweiler, D., & Crain, S. (1986). Language mechanisms and reading disorder. *Cognition*, 24, 139–168.
- Siegler, R. S. (1988). Individual differences in strategy choices: Good students, not-so-good students, and perfectionists. *Child Development*, 59, 833–851.
- Singer, J. D., (2002). Fitting individual growth models using SAS PROC MIXED. In Moskowitz & S. Hersberger (Eds.), *Modeling intraindividual variability with repeated measures data: Methods and applications* (pp. 135–170). Mahwah, NJ: Erlbaum.
- Snijders, T. A., & Bosker, R. J. (2003). *Multilevel analysis*. Thousand Oaks, CA: Sage.
- Stoner, S. B. (1982). Age differences in crystallized and fluid intellectual abilities. *Journal of Psychology*, 110, 7–10.
- Swanson, H. L. (1992). Generality and modifiability of working memory among skilled and less skilled readers. *Journal of Educational Psychology*, 84, 473–488.
- Swanson, H. L. (1995). *S-Cognitive Processing Test*. Austin, TX: PRO-ED.
- Swanson, H. L. (1999). Reading comprehension and working memory in skilled readers: Is the phonological loop more important than the executive system? *Journal of Experimental Child Psychology*, 72, 1–31.
- Swanson, H. L. (2003). Age-related differences in learning disabled and skilled readers' working memory. *Journal of Experimental Child Psychology*, 85, 1–31.
- Swanson, H. L. (2004). Working memory and phonological processing as predictors of children's mathematical problem solving at different ages. *Memory & Cognition*, 32, 648–661.
- Swanson, H. L. (2006). Cross-sectional and incremental changes in working memory and mathematical problem solving. *Journal of Educational Psychology*, 98, 265–281.
- Swanson, H. L., & Ashbaker, M. (2000). Working memory, short-term memory, articulation speed, word recognition, and reading comprehension in learning disabled readers: Executive and/or articulatory system? *Intelligence*, 28, 1–30.
- Swanson, H. L., Ashbaker, M., & Lee, C. (1996). Learning disabled readers' working memory as a function of processing demands. *Journal of Experimental Child Psychology*, 61, 242–275.
- Swanson, H. L., & Beebe-Frankenberger, M. (2004). The relationship between working memory and mathematical problem solving in children at risk and not a risk for serious math difficulties. *Journal of Educational Psychology*, 96, 471–491.
- Swanson, H. L., & Berninger, V. (1995). The role of working memory and STM in skilled and less skilled readers' word recognition and comprehension. *Intelligence*, 21, 83–108.
- Swanson, H. L., Cooney, J. B., & Brock, S. (1993). The influence of working memory and classification ability on children's word problem solution. *Journal of Experimental Child Psychology*, 55, 374–395.
- Swanson, H. L., & Jerman, O. (2006). Math disabilities: A selective meta-analysis of the literature. *Review of Educational Research*, 76, 249–274.
- Swanson, H. L., Mink, J., & Bocian, K. M. (1999). Cognitive processing deficits in poor readers with symptoms of reading disabilities and ADHD: More alike than different? *Journal of Educational Psychology*, 91, 321–353.
- Swanson, H. L., & Sachse-Lee, C. (2001). Mathematic problem solving and working memory in children with learning disabilities: Both exec-

- utive and phonological processes are important. *Journal of Experimental Child Psychology*, 79, 294–321.
- Towse, J. (1998). On random generation and the central executive of working memory. *British Journal of Psychology*, 89, 77–101.
- Wagner, R., & Torgesen, J. (1999). *Test of Word Reading Efficiency*. Austin, TX: PRO-ED.
- Wagner, R., Torgesen, J., & Rashotte, C. (2000). *Comprehensive Test of Phonological Processes*. Austin, TX: PRO-ED.
- Whitney, P., Arnett, P., Driver, A., & Budd, D. (2001). Measuring central executive functioning: What's in a reading span? *Brain & Cognition*, 45, 1–14.
- Wilkinson, G. S. (1993). *The Wide Range Achievement Test—Third Edition*. Wilmington, DE: Wide Range.
- Wilson, R. S., Beckett, L. A., Barnes, L., Schneider, J., Bach, J., Evans, D., & Bennett, D. A. (2002). Individual differences in rates of change in cognitive abilities of older persons. *Psychology and Aging*, 17, 179–193.
- Woodcock, R. W. (1998). *Woodcock Reading Mastery Test—Revised (Form G)*. Circle Pines, MN: American Guidance

## Appendix A

Variable, Sample Size, Means, and Standard Deviations for Wave 1  
(Grades 1, 2, and 3)

Variable	At risk for SMD			Not at risk for SMD		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
Chronological age (in months)						
Age 1	77	80.09	4.55	54	80.04	5.38
Age 2	69	92.64	4.70	45	92.84	5.35
Age 3	64	105.42	7.69	44	104.09	4.85
Fluid intelligence (Raven)						
RAV-SS 1	77	102.29	15.99	54	111.07	12.78
RAV-SS 2	69	106.68	12.86	45	115.00	13.04
RAV-SS 3	64	105.42	14.20	44	118.89	10.86
RAV-R 1	77	18.62	5.30	54	22.06	5.39
RAV-R 2	69	22.96	5.40	45	26.53	5.67
RAV-R 3	64	25.69	5.75	44	30.57	3.91
Word problems (WISC-III)						
Arithm-SC 1	77	6.95	.80	54	12.94	2.22
Arithm-SC 2	69	8.65	1.53	45	11.02	3.43
Arithm-SC 3	64	8.92	1.86	44	11.18	2.37
Arithm-R 1	77	8.65	2.68	54	12.80	1.02
Arithm-R 2	69	11.83	2.20	45	13.36	2.13
Arithm-R 3	64	13.47	2.24	44	15.09	1.67
Digit naming speed (CTOPP)						
RDN-SC 1	77	6.99	2.72	54	8.51	1.61
RDN-SC 2	69	8.65	1.86	45	10.04	2.31
RDN-SC 3	64	8.92	3.09	44	11.38	2.47
RDN-R 1	77	59.31	14.75	54	43.72	8.95
RDN-R 2	69	44.54	8.80	45	36.47	8.76
RDN-R 3	64	37.83	7.52	44	31.93	7.62
Word problems: Semantic variation						
WPSV 1	77	2.06	1.66	54	3.04	1.61
WPSV 2	69	3.16	1.49	45	3.49	1.08
WPSV 3	64	3.78	1.33	44	4.09	1.93
Word problem solving processes (total of problem solving components)						
COMPON 1	77	6.77	2.95	54	8.67	2.56
COMPON 2	69	8.58	3.34	46	10.09	3.00
COMPON 3	64	10.39	3.03	43	12.60	2.03
Knowledge of problem solving components						
QUES 1	77	1.47	0.90	54	2.15	0.81
QUES 2	69	1.87	0.95	46	2.26	0.93
QUES 3	64	2.22	0.83	43	2.65	0.65
NUMB 1	77	1.53	1.06	54	2.15	0.98
NUMB 2	69	2.09	0.98	46	2.28	0.81
NUMB 3	64	2.56	0.73	43	2.81	0.45
Goal 1	77	1.19	0.84	54	1.15	0.92
Goal 2	69	0.97	0.86	46	1.39	0.95
Goal 3	64	1.81	0.97	43	2.40	0.69
OPER 1	77	1.44	0.75	54	1.61	0.81
OPER 2	69	1.90	0.99	46	2.15	0.82

(Appendixes continue)



Appendix A (*continued*)

Variable	At risk for SMD			Not at risk for SMD		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
OPER 3	64	1.58	1.12	43	2.09	1.02
ALGO 1	77	1.13	0.83	54	1.61	0.86
ALGO 2	69	1.75	0.95	46	2.00	0.79
ALGO 3	64	2.22	0.86	43	2.65	0.53
IRRE 1	77	1.86	0.81	54	2.00	0.73
IRRE 2	69	2.04	0.96	46	2.24	0.79
IRRE 3	64	2.45	0.85	43	2.86	0.41
Math calculation (WRAT, WIAT, CBM)						
AWRAT-SS 1	77	111.27	11.14	54	115.35	10.37
AWRAT-SS 2	69	101.17	11.36	45	106.29	9.93
AWRAT-SS 3	64	106.86	14.79	44	116.18	12.55
AWRAT-R 1	77	18.77	1.59	54	19.69	1.49
AWRAT-R 2	69	21.86	2.14	45	23.07	1.96
AWRAT-R 3	64	27.45	3.74	44	29.34	3.23
WIAT-SS 1	77	103.23	13.87	54	110.19	12.95
WIAT-SS 2	69	102.72	14.78	45	107.80	12.24
WIAT-SS 3	62	109.69	14.17	43	118.23	10.63
WIAT-R 1	77	8.44	2.30	54	10.07	1.87
WIAT-R 2	69	13.87	3.19	45	15.36	2.83
WIAT-R 3	62	19.63	3.59	43	22.05	2.54
CBM-R 1	77	26.81	12.03	54	28.39	11.25
CBM-R 2	69	26.48	11.59	45	27.82	10.00
CBM-R 3	64	25.95	16.12	43	30.79	12.73
Phonological knowledge (Elision CTOPP, Pseudowords TOWRE)						
EL-SC 1	77	8.88	3.23	54	12.13	3.29
EL-SC 2	69	9.17	3.21	45	11.27	3.16
EL-SC 3	64	9.03	3.09	44	11.09	2.94
EL-R 1	77	5.48	4.08	54	10.24	4.90
EL-R 2	69	8.87	4.79	45	12.24	4.90
EL-R 3	64	10.66	4.92	44	13.91	4.66
PW-SS 1	77	101.22	13.10	54	112.93	12.81
PW-SS 2	69	101.45	12.64	45	112.07	12.66
PW-SS 3	64	101.06	13.61	44	113.05	15.69
PW-R 1	77	9.78	8.04	54	18.69	10.18
PW-R 2	69	18.96	10.33	45	28.51	11.00
PW-R 3	64	25.23	12.06	44	35.18	12.33
Reading skill (WRAT-III, Real Words TOWRE, Comprehension WRMT)						
WREAD-SS 1	77	103.64	16.65	54	116.50	15.05
WREAD-SS 2	69	100.57	16.66	45	107.91	12.23
WREAD-SS 3	64	99.16	16.90	44	110.23	14.52
WREAD-R 1	77	21.23	4.60	54	25.61	4.24
WREAD-R 2	69	27.32	5.17	45	29.91	3.72
WREAD-R 3	64	30.08	5.87	44	34.07	5.25
SWE-SS 1	77	100.35	12.77	54	113.78	12.43
SWE-SS 2	69	101.58	17.48	45	115.31	12.62
SWE-SS 3	64	101.38	14.51	44	113.36	12.22
SWE-TC 1	77	21.94	13.24	54	38.33	15.14
SWE-TC 2	69	44.42	17.64	45	58.53	12.59
SWE-TC 3	64	53.17	17.09	44	66.30	11.61
WRMT-SS 1	77	101.77	13.21	54	112.22	10.45
WRMT-SS 2	69	101.45	14.20	45	110.36	8.77
WRMT-SS 3	64	98.08	10.79	44	107.91	8.49
WRMT-R 1	77	11.71	7.97	54	20.28	7.76
WRMT-R 2	69	21.94	10.10	45	29.24	6.19
WRMT-R 3	64	26.95	8.72	44	35.39	5.04
Letter naming speed (CTOPP)						
RLN-SC 1	77	9.09	1.60	54	10.85	1.57
RLN-SC 2	69	10.20	1.87	45	11.62	2.27
RLN-SC 3	64	10.77	3.88	44	11.59	3.07
RLN-R 1	77	72.09	23.90	54	49.33	9.51
RLN-R 2	69	51.68	11.80	45	41.38	9.06
RLN-R 3	64	43.95	10.08	44	37.23	9.00
Short-term memory (phonological loop)						
RWORD 1	77	6.61	2.88	54	8.37	3.28

Appendix A (*continued*)

Variable	At risk for SMD			Not at risk for SMD		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
RWORD 2	69	7.48	3.00	45	7.84	3.74
RWORD 3	64	7.70	3.12	44	9.00	3.38
PWORD 1	77	2.68	1.63	54	3.44	1.93
PWORD 2	69	2.45	1.41	45	2.76	1.48
PWORD 3	64	2.72	1.47	44	2.91	1.64
DIG-F 1	77	4.62	2.41	54	5.69	3.02
DIG-F 2	69	5.00	2.56	45	5.49	3.11
DIG-F 3	64	5.52	2.85	44	7.55	2.91
DIG-B 1	77	2.32	0.95	54	3.00	1.06
DIG-B 2	69	2.81	1.34	45	3.62	1.54
DIG-B 3	64	3.28	1.24	44	4.18	1.67
Working memory (executive)						
LISSPAN 1	77	2.51	3.08	54	4.04	5.17
LISSPAN 2	69	3.67	3.69	45	6.34	5.18
LISSPAN 3	61	6.88	5.01	43	9.56	5.69
SEMSET 1	77	3.62	2.88	54	4.63	3.56
SEMSET 2	69	5.35	5.15	45	8.49	6.39
SEMSET 3	64	6.06	4.61	44	7.66	6.29
AUDSET 1	77	3.79	3.25	54	4.48	3.60
AUDSET 2	69	5.75	4.62	45	4.58	3.96
AUDSET 3	64	7.44	5.38	44	11.91	7.38
UPDATE 1	77	2.61	3.53	54	5.37	4.89
UPDATE 2	64	3.81	3.67	44	6.11	4.45
UPDATE 3	64	4.03	3.75	45	6.09	4.32
Working memory: Visual-spatial						
MATRIX 1	77	4.94	3.39	54	5.31	3.26
MATRIX 2	69	5.97	3.01	45	7.24	3.20
MATRIX 3	64	7.52	3.15	44	8.93	4.27
MAP-DIR 1	77	3.77	2.63	54	3.26	1.67
MAP-DIR 2	69	2.33	2.49	45	3.76	4.73
MAP-DIR 3	64	3.19	3.80	44	5.64	5.72
Fluency						
CATF 1	77	9.66	4.19	54	12.59	4.46
CATF 2	69	11.19	4.93	45	14.42	4.50
CATF 3	64	13.00	4.12	44	15.18	4.76
LETf 1	77	5.47	3.02	54	6.59	3.67
LETf 2	69	6.72	3.46	45	7.76	3.28
LETf 3	64	7.47	3.47	44	9.16	3.65
Random generation						
RANDL 1	77	0.49	0.30	54	0.42	0.24
RANDL 2	67	0.51	0.28	45	0.44	0.23
RANDL 3	64	0.40	0.19	44	0.37	0.15
RANDN 1	77	0.23	0.18	54	0.24	0.17
RANDN 2	69	0.26	0.13	45	0.27	0.12
RANDN 3	64	0.27	0.52	44	0.23	0.09

*Note.* The numbers 1, 2, and 3 at the end of each abbreviation refer to Waves 1, 2, and 3, respectively. SC = scale score ( $M = 10$ ,  $SD = 2$ ); SS = standard score; R = Raw score; RAV = Raven Progressive Matrices Test; WISC-III = Wechsler Intelligence Scale for Children—Third Edition; Arithm = word problem for the WISC-III; CTOPP = Comprehensive Test of Phonological Processing; RDN = rapid digit naming; WPSV = word problems with semantic variations; COMPON = total of knowledge of word problem components; QUES = question; NUMB = number; OPER = operations; ALGO = algorithm knowledge; IRRE = Irrelevant; WRAT = Wide Range Achievement Test; WIAT = Wechsler Individual Achievement Test; CBM = Curriculum; AWRAT = arithmetic subtest of WRAT; TWORE = Test of Word Reading Efficiency; EL = Elision subtest (CTOPP); PW = pseudoword fluency (TOWRE); WRAT-III = Wide Range Achievement Test—Third Edition; WREAD = WRAT reading score; SWE = Real word fluency (TOWRE); WRMT = passage comprehension Woodcock Reading Mastery Test—Revised; RLN = rapid naming of letters; RWORD = real word recall; PWORD = pseudoword recall; DIG-F = digits forward task; DIG-B = digits backward task; LISSPAN = Listening Span task; SEMSET = Semantic Association Span task; AUDSET = auditory Digit/Sentence Span task; UPDATE = Updating Task; MATRIX = visual-spatial matrix task; MAP-DIR = mapping and direction task; CATF = categorical fluency; LETf = letter fluency; RANDL = random generation of letters; RANDN = random generation of numbers.

(Appendixes continue)



## Appendix B

## Variable, Sample Size, Means, and Standard Deviations Starting at Grade 2

Variable	At risk for SMD			Not at risk for SMD		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
Chronological age (in months)						
Age 1	34	94.06	5.25	58	92.79	3.88
Age 2	32	106.13	4.95	53	105.43	4.16
Age 3	30	117.80	4.82	53	117.15	5.20
Fluid intelligence (Raven)						
RAV-SS 1	34	104.18	11.33	58	111.98	15.91
RAV-SS 2	32	105.91	12.78	53	112.74	12.67
RAV-SS 3	30	107.27	10.93	53	111.91	11.23
RAV-R 1	34	23.41	4.86	58	25.86	5.83
RAV-R 2	32	26.50	5.18	53	28.98	4.63
RAV-R 3	30	29.40	3.87	53	30.72	3.81
Word problems (WISC-III)						
Arithm-SC 1	33	8.00	1.80	58	11.76	2.18
Arithm-SC 2	32	9.09	3.11	53	11.74	2.59
Arithm-SC 3	30	8.97	2.77	53	11.02	2.68
Arithm-R 1	33	11.79	1.73	58	14.19	1.19
Arithm-R 2	32	13.81	2.09	53	15.74	1.84
Arithm-R 3	30	15.10	1.99	53	16.66	2.08
Digit naming speed (CTOPP)						
RDN-SC 1	33	7.78	1.01	58	9.35	1.53
RDN-SC 2	32	8.84	1.64	53	10.82	2.00
RDN-SC 3	30	10.06	2.04	53	11.86	2.16
RDN-R 1	33	48.76	10.62	58	37.91	6.33
RDN-R 2	32	41.91	10.13	53	33.17	6.47
RDN-R 3	30	36.53	7.66	53	30.40	6.22
Word problems: Semantic variation						
WPSV 1	33	3.21	1.24	58	3.78	1.23
WPSV 2	32	3.75	1.02	53	3.96	1.19
WPSV 3	30	4.17	1.26	53	4.51	1.61
Word problem solving processes (total of problem solving components)						
COMPON 1	34	8.38	3.08	58	10.47	2.54
COMPON 2	32	11.53	2.14	53	12.79	2.01
COMPON 3	30	16.40	2.14	53	17.60	2.61
Knowledge of problem solving components						
QUES 1	34	1.62	0.85	58	1.98	0.87
QUES 2	32	2.28	1.02	53	2.62	0.60
QUES 3	30	3.23	1.04	53	3.58	0.66
NUMB 1	34	2.06	0.89	58	2.43	0.68
NUMB 2	32	2.50	0.62	53	2.79	0.45
NUMB 3	30	3.50	0.63	53	3.75	0.52
Goal 1	34	0.85	0.86	58	1.45	0.98
Goal 2	32	2.09	0.82	53	2.28	0.91
Goal 3	30	3.07	0.83	53	3.25	0.98
OPER 1	34	1.91	1.03	58	2.41	0.77
OPER 2	32	2.19	0.54	53	2.43	0.69
OPER 3	30	3.13	0.51	53	3.36	0.88
ALGO 1	34	1.94	0.92	58	2.19	0.71
ALGO 2	32	2.47	0.62	53	2.66	0.59
ALGO 3	30	3.47	0.63	53	3.66	0.62
IRRE 1	34	2.03	0.90	58	2.40	0.70
IRRE 2	32	2.59	0.71	53	2.79	0.53
IRRE 3	30	2.83	0.46	53	2.70	0.61
Math calculation (WRAT, WIAT, CBM)						
AWRAT-SS 1	34	99.59	9.97	58	106.12	9.62
AWRAT-SS 2	32	107.00	16.13	53	118.55	13.26
AWRAT-SS 3	30	106.93	14.29	53	116.49	10.61
AWRAT-R 1	34	21.94	1.72	58	23.34	1.87
AWRAT-R 2	32	27.94	4.14	53	31.00	3.24
AWRAT-R 3	30	31.37	4.06	53	34.13	2.76
WIAT-SS 1	34	102.06	12.13	58	108.86	10.55
WIAT-SS 2	32	111.13	15.50	53	119.96	13.45
WIAT-SS 3	30	108.23	15.49	53	117.36	11.32

## Appendix B (continued)

Variable	At risk for SMD			Not at risk for SMD		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
WIAT-R 1	34	14.29	2.74	58	15.72	2.14
WIAT-R 2	32	20.56	3.75	53	22.94	3.17
WIAT-R 3	30	23.97	4.79	53	26.77	2.94
CBM-R 1	34	27.35	18.51	58	27.48	11.26
CBM-R 2	32	37.75	17.52	53	33.36	11.60
CBM-R 3	30	70.10	20.45	53	75.66	26.23
Phonological knowledge (Elision CTOPP, Pseudowords TOWRE)						
EL-SC 1	33	9.45	3.34	58	11.17	3.15
EL-SC 2	32	8.94	2.38	53	10.64	3.16
EL-SC 3	30	8.90	2.83	53	11.30	2.95
EL-R 1	33	9.82	5.29	58	12.28	4.89
EL-R 2	32	10.84	4.06	53	13.28	4.92
EL-R 3	30	12.10	4.57	53	15.68	4.24
PW-SS 1	33	101.45	11.73	58	111.26	12.47
PW-SS 2	32	100.84	12.49	53	111.13	13.59
PW-SS 3	30	97.23	13.04	53	106.68	12.25
PW-R 1	33	18.70	9.82	58	27.48	10.86
PW-R 2	32	24.59	10.73	53	33.60	10.62
PW-R 3	30	26.77	11.88	53	35.45	10.26
Reading skill (WRAT-III, Real Words TOWRE, Comprehension WRMT)						
WREAD-SS 1	33	98.70	12.91	58	108.41	12.83
WREAD-SS 2	32	97.31	12.85	53	106.83	13.23
WREAD-SS 3	30	99.83	15.03	53	110.00	12.85
WREAD-R 1	33	27.36	4.22	58	30.76	4.27
WREAD-R 2	32	29.88	4.19	53	33.25	4.51
WREAD-R 3	30	33.13	5.14	53	36.85	4.40
SWE-SS 1	33	103.18	13.34	58	115.47	12.16
SWE-SS 2	32	102.97	11.33	53	113.13	10.12
SWE-SS 3	30	99.57	11.74	53	108.96	10.36
SWE-TC 1	33	44.42	13.21	58	58.12	10.40
SWE-TC 2	32	55.94	12.18	53	66.25	8.95
SWE-TC 3	30	62.17	12.65	53	71.32	9.39
WRMT-SS 1	33	103.03	10.24	58	109.07	10.70
WRMT-SS 2	32	98.84	8.69	53	106.87	9.66
WRMT-SS 3	30	97.60	7.41	53	104.57	10.55
WRMT-R 1	33	25.00	6.60	58	29.84	5.96
WRMT-R 2	32	29.34	5.94	53	34.45	5.66
WRMT-R 3	30	33.47	4.85	53	37.49	6.32
Letter naming speed (CTOPP)						
RLN-SC 1	33	9.33	1.80	58	10.93	1.99
RLN-SC 2	32	9.47	2.00	53	10.79	2.11
RLN-SC 3	30	9.70	2.37	53	10.75	2.03
RLN-R 1	33	52.52	12.69	58	42.88	7.97
RLN-R 2	32	45.78	10.62	53	39.04	8.60
RLN-R 3	30	39.80	8.81	53	35.06	6.52
Short-term memory (phonological loop)						
RWORD 1	33	8.09	3.51	58	9.53	3.15
RWORD 2	32	7.59	3.66	53	9.60	3.41
RWORD 3	30	8.50	2.96	53	10.66	3.51
PWORD 1	33	3.70	1.85	58	3.90	2.24
PWORD 2	32	2.59	1.78	53	3.21	1.79
PWORD 3	30	3.23	1.91	53	2.98	1.74
DIG-F 1	33	6.03	2.82	58	6.84	2.73
DIG-F 2	32	5.69	3.01	53	7.09	3.36
DIG-F 3	30	6.20	3.27	53	8.34	2.59
DIG-B 1	33	2.82	0.98	58	3.43	1.30
DIG-B 2	32	3.34	1.23	53	4.13	1.87
DIG-B 3	30	3.50	1.46	53	4.66	2.39

(Appendixes continue)



## Appendix B (continued)

Variable	At risk for SMD			Not at risk for SMD		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
Working memory (executive)						
LISSPAN 1	33	4.88	4.04	58	6.59	4.95
LISSPAN 2	31	5.63	4.89	52	8.65	4.62
LISSPAN 3	30	8.25	4.04	52	10.36	4.64
SEMSET 1	33	5.27	4.22	58	4.60	3.55
SEMSET 2	32	6.09	5.11	51	7.88	6.28
SEMSET 3	30	7.57	5.83	53	9.08	7.98
AUDSET 1	33	4.64	5.02	58	5.45	4.74
AUDSET 2	32	6.75	4.20	53	11.11	7.10
AUDSET 3	30	10.87	6.98	53	13.81	9.15
UPDATE 1	33	3.94	3.66	58	7.00	5.06
UPDATE 2	30	4.47	3.78	53	6.57	4.58
UPDATE 3	30	4.60	3.77	51	6.98	4.92
Working memory: Visual-spatial						
MATRIX 1	34	5.97	4.12	58	6.83	3.74
MATRIX 2	32	6.84	3.33	53	7.75	4.55
MATRIX 3	30	8.47	4.20	53	11.13	5.95
MAP-DIR 1	33	3.42	4.36	58	3.76	2.38
MAP-DIR 2	32	4.25	5.10	53	3.94	4.80
MAP-DIR 3	30	7.33	6.91	53	6.74	6.29
Fluency						
CATF 1	33	12.76	3.98	58	13.86	3.96
CATF 2	32	13.03	3.94	53	15.64	4.68
CATF 3	30	14.50	3.91	53	16.42	4.47
LETF 1	33	7.09	2.79	58	7.07	3.09
LETF 2	32	7.56	2.83	53	8.21	3.26
LETF 3	30	8.43	2.85	53	9.21	4.78
Random generation						
RANDL 1	34	0.48	0.27	58	0.45	0.22
RANDL 2	32	0.50	0.23	53	0.48	0.20
RANDL 3	30	0.39	0.16	53	0.36	0.12
RANDN 1	34	0.15	0.10	58	0.23	0.15
RANDN 2	32	0.24	0.22	53	0.24	0.11
RANDN 3	30	0.23	0.11	53	0.26	0.18

*Note.* The numbers 1, 2, and 3 at the end of each abbreviation refer to Waves 1, 2, and 3, respectively. SC = scale score ( $M = 10$ ,  $SD = 2$ ); SS = standard score; R = Raw score; RAV = Raven Progressive Matrices Test; WISC-III = Wechsler Intelligence Scale for Children—Third Edition; Arithm = word problem for the WISC-III; CTOPP = Comprehensive Test of Phonological Processing; RDN = rapid digit naming; WPSV = word problems with semantic variations; COMPON = total of knowledge of word problem components; QUES = question; NUMB = number; OPER = operations; ALGO = algorithm knowledge; IRRE = Irrelevant; WRAT = Wide Range Achievement Test; WIAT = Wechsler Individual Achievement Test; CBM = Curriculum; AWRA = arithmetic subtest of WRAT; TWORE = Test of Word Reading Efficiency; EL = Elision subtest (CTOPP); PW = pseudoword fluency (TOWRE); WRAT-III = Wide Range Achievement Test—Third Edition; WREAD = WRAT reading score; SWE = Real word fluency (TOWRE); WRMT = passage comprehension Woodcock Reading Mastery Test—Revised; RLN = rapid naming of letters; RWORD = real word recall; PWORD = pseudoword recall; DIG-F = digits forward task; DIG-B = digits backward task; LISSPAN = Listening Span task; SEMSET = Semantic Association Span task; AUDSET = auditory Digit/Sentence Span task; UPDATE = Updating Task; MATRIX = visual-spatial matrix task; MAP-DIR = mapping and direction task; CATF = categorical fluency; LETF = letter fluency; RANDL = random generation of letters; RANDN = random generation of numbers.

## Appendix C

Variable, Sample Size, Means, and Standard Deviations Starting at Grade 3

Variable	At risk for SMD			Not at risk for SMD		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
Chronological age (in months)						
Age 1	24	107.88	6.35	107	104.37	4.91
Age 2	22	120.82	6.61	98	116.60	4.09
Age 3	22	132.73	6.27	89	128.18	6.71
Fluid intelligence (Raven)						
RAV-SS 1	24	96.71	9.70	107	108.47	13.42
RAV-SS 2	22	99.95	10.96	98	110.36	11.57
RAV-SS 3	22	102.27	10.17	88	109.07	13.81
RAV-R 1	24	23.79	4.55	107	27.67	5.00
RAV-R 2	22	27.00	4.35	98	30.09	4.04
RAV-R 3	22	29.95	3.54	88	31.25	3.95
Word problems (WISC-III)						
Arithm-SC 1	22	6.09	1.95	107	11.70	2.08
Arithm-SC 2	22	8.73	3.34	98	10.77	3.07
Arithm-SC 3	21	7.05	1.94	89	10.80	3.49
Arithm-R 1	22	11.91	2.51	107	15.74	1.57
Arithm-R 2	22	14.86	1.52	98	16.46	2.10
Arithm-R 3	21	14.90	1.61	89	17.79	3.26
Digit naming speed (CTOPP)						
RDN-SC 1	22	9.25	1.44	106	11.03	2.24
RDN-SC 2	22	10.77	2.10	98	12.04	2.49
RDN-SC 3	21	11.83	2.55	89	12.68	2.21
RDN-R 1	22	38.73	7.08	107	32.66	6.84
RDN-R 2	22	33.55	6.51	98	29.97	6.92
RDN-R 3	21	30.48	6.96	89	28.09	5.99
Word problems: Semantic variation						
WPSV 1	22	3.23	1.57	107	3.68	1.29
WPSV 2	22	3.73	1.49	98	4.18	1.12
WPSV 3	21	3.71	0.72	89	4.17	1.36
Word problem solving processes (total of problem solving components)						
COMPON 1	24	9.54	2.98	107	11.91	2.58
COMPON 2	22	15.05	2.61	98	17.29	2.43
COMPON 3	22	20.05	2.61	89	22.35	2.50
Knowledge of problem solving components						
QUES 1	24	1.79	1.06	107	2.45	0.78
QUES 2	22	2.82	0.80	98	3.47	0.71
QUES 3	22	3.82	0.80	89	4.48	0.71
NUMB 1	24	2.58	0.58	107	2.61	0.59
NUMB 2	22	3.45	0.67	98	3.56	0.59
NUMB 3	22	4.45	0.67	89	4.55	0.60
Goal 1	24	1.25	0.99	107	2.19	0.92
Goal 2	22	2.64	0.79	98	3.16	0.77
Goal 3	22	3.64	0.79	89	4.18	0.79
OPER 1	24	1.92	0.97	107	2.30	0.84
OPER 2	22	3.05	0.84	98	3.50	0.71
OPER 3	22	4.05	0.84	89	4.53	0.69
ALGO 1	24	2.00	0.88	107	2.36	0.73
ALGO 2	22	3.09	0.87	98	3.59	0.66
ALGO 3	22	4.09	0.87	89	4.61	0.65
IRRE 1	24	2.50	0.66	107	2.63	0.64
IRRE 2	22	2.73	0.55	98	2.86	0.35
IRRE 3	22	2.36	0.90	88	2.73	0.56
Math calculation (WRAT, WIAT, CBM)						
AWRAT-SS 1	24	106.79	14.83	107	117.61	12.09
AWRAT-SS 2	22	105.77	11.99	98	117.88	9.45
AWRAT-SS 3	22	99.36	13.59	88	113.72	11.10

(Appendixes continue)



## Appendix C (continued)

Variable	At risk for SMD			Not at risk for SMD		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
AWRAT-R 1	24	28.25	3.64	107	30.13	2.89
AWRAT-R 2	22	31.68	3.14	98	34.27	2.43
AWRAT-R 3	22	32.77	3.66	88	36.32	3.54
WIAT-SS 1	24	108.17	13.74	107	116.63	9.94
WIAT-SS 2	22	108.32	12.63	98	119.68	9.88
WIAT-SS 3	22	98.41	14.70	88	111.34	10.04
WIAT-R 1	24	20.17	2.90	107	21.66	2.11
WIAT-R 2	22	25.05	3.30	98	27.35	2.81
WIAT-R 3	22	25.86	3.89	88	28.88	3.10
CBM-R 1	24	31.75	11.48	107	35.18	12.87
CBM-R 2	22	79.05	53.72	98	82.18	28.60
CBM-R 3	22	88.86	63.09	88	90.34	61.69
Phonological knowledge (Elision CTOPP, Pseudowords TOWRE)						
EL-SC 1	22	7.86	4.14	107	10.85	3.52
EL-SC 2	22	7.68	3.17	98	10.79	2.90
EL-SC 3	21	8.14	3.32	89	10.64	2.66
EL-R 1	22	9.59	5.74	107	13.53	5.41
EL-R 2	22	10.68	4.63	98	15.00	4.34
EL-R 3	21	13.14	4.92	89	16.25	3.84
PW-SS 1	22	96.18	11.16	107	110.74	13.42
PW-SS 2	22	99.64	13.83	98	108.74	14.22
PW-SS 3	21	93.38	13.23	89	104.72	13.41
PW-R 1	22	22.27	10.13	107	33.63	10.59
PW-R 2	22	29.27	12.57	98	36.90	11.43
PW-R 3	21	30.76	12.16	89	40.11	10.48
Reading skill (WRAT-III, Real Words TOWRE, Comprehension WRMT)						
WREAD-SS 1	22	94.32	16.70	107	107.18	11.51
WREAD-SS 2	22	94.68	16.46	98	107.39	12.16
WREAD-SS 3	21	91.81	16.78	89	108.94	12.71
WREAD-R 1	22	29.59	5.34	107	33.41	4.10
WREAD-R 2	22	32.09	4.97	98	35.74	4.37
WREAD-R 3	21	33.43	5.64	89	38.89	4.67
SWE-SS 1	22	99.23	14.72	107	112.29	10.38
SWE-SS 2	22	99.55	11.68	98	109.65	10.84
SWE-SS 3	21	94.24	11.57	89	105.53	9.92
SWE-TC 1	22	53.64	15.77	107	65.79	8.99
SWE-TC 2	22	61.68	13.05	98	71.68	10.09
SWE-TC 3	21	64.86	12.63	89	76.20	9.19
WRMT-SS 1	22	95.32	12.42	107	105.87	8.90
WRMT-SS 2	22	90.95	9.45	98	103.93	9.40
WRMT-SS 3	21	90.29	9.50	89	102.18	10.51
WRMT-R 1	22	28.14	8.20	107	34.27	4.60
WRMT-R 2	22	29.32	5.49	98	36.92	5.29
WRMT-R 3	21	33.67	5.76	89	40.09	6.44
Letter naming speed (CTOPP)						
RLN-SC 1	22	9.64	1.92	106	11.10	2.08
RLN-SC 2	22	9.77	2.51	98	11.10	2.65
RLN-SC 3	21	9.90	3.10	89	11.37	4.11
RLN-R 1	22	42.77	9.00	106	37.93	7.21
RLN-R 2	22	38.09	9.40	98	35.03	7.34
RLN-R 3	21	35.05	10.62	89	32.54	7.24
Short-term memory (phonological loop)						
RWORD 1	22	7.86	2.95	107	9.16	3.08
RWORD 2	22	7.45	3.13	98	10.03	3.81
RWORD 3	21	7.71	2.43	88	10.88	3.63
PWORD 1	22	2.95	1.62	107	3.82	1.96
PWORD 2	22	2.59	1.37	98	3.47	1.93
PWORD 3	21	2.95	1.69	89	3.51	1.61
DIG-F 1	22	4.68	2.12	107	7.22	3.09
DIG-F 2	22	5.86	2.71	98	6.94	3.00
DIG-F 3	21	7.29	2.53	89	8.20	3.00
DIG-B 1	22	2.82	1.05	107	3.61	1.61
DIG-B 2	22	3.50	1.74	98	4.07	1.83
DIG-B 3	21	3.62	1.28	89	4.62	1.78

## Appendix C (continued)

Variable	At risk for SMD			Not at risk for SMD		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
Working memory (executive)						
LISSPAN 1	22	4.02	4.25	107	6.77	4.06
LISSPAN 2	22	4.73	5.80	94	8.94	4.86
LISSPAN 3	21	7.60	5.40	86	11.31	5.08
SEMSET 1	22	4.64	3.55	106	6.54	5.34
SEMSET 2	22	7.23	6.02	97	9.68	7.11
SEMSET 3	21	6.86	4.56	88	11.86	8.81
AUDSET 1	22	4.68	3.85	107	7.87	6.13
AUDSET 2	22	9.27	5.36	98	12.88	7.94
AUDSET 3	21	11.71	6.68	89	14.61	8.92
UPDATE 1	22	2.95	3.24	107	8.24	5.19
UPDATE 2	21	5.76	4.36	89	7.78	4.42
UPDATE 3	20	5.50	4.30	91	7.44	4.40
Working memory: Visual-spatial						
MATRIX 1	24	7.04	3.61	107	7.93	4.65
MATRIX 2	22	8.50	4.01	98	9.20	5.63
MATRIX 3	22	9.45	5.35	88	10.97	6.28
MAP-DIR 1	22	3.91	3.08	107	5.05	4.63
MAP-DIR 2	22	4.09	4.77	97	5.66	6.54
MAP-DIR 3	21	4.81	5.36	89	7.34	6.89
Fluency						
CATF 1	22	13.23	3.98	107	15.35	4.27
CATF 2	22	14.09	4.77	98	16.88	5.24
CATF 3	21	14.43	4.08	89	16.92	5.17
LETF 1	22	7.05	2.89	107	8.39	3.39
LETF 2	22	9.36	5.09	98	8.88	3.47
LETF 3	21	8.76	4.00	89	10.25	3.48
Random generation						
RANDL 1	24	0.54	0.25	107	0.40	0.18
RANDL 2	22	0.47	0.21	98	0.38	0.19
RANDL 3	22	0.37	0.09	88	0.33	0.13
RANDN 1	24	0.17	0.13	107	0.20	0.11
RANDN 2	22	0.19	0.08	98	0.22	0.09
RANDN 3	22	0.19	0.07	88	0.20	0.11

*Note.* The numbers 1, 2, and 3 at the end of each abbreviation refer to Waves 1, 2, and 3, respectively. SC = scale score ( $M = 10$ ,  $SD = 2$ ); SS = standard score; R = Raw score; RAV = Raven Progressive Matrices Test; WISC-III = Wechsler Intelligence Scale for Children—Third Edition; Arithm = word problem for the WISC-III; CTOPP = Comprehensive Test of Phonological Processing; RDN = rapid digit naming; WPSV = word problems with semantic variations; COMON = total of knowledge of word problem components; QUES = question; NUMB = number; OPER = operations; ALGO = algorithm knowledge; IRRE = Irrelevant; WRAT = Wide Range Achievement Test; WIAT = Wechsler Individual Achievement Test; CBM = Curriculum; AWRAT = arithmetic subtest of WRAT; TWORE = Test of Word Reading Efficiency; EL = Elision subtest (CTOPP); PW = pseudoword fluency (TOWRE); WRAT-III = Wide Range Achievement Test—Third Edition; WREAD = WRAT reading score; SWE = Real word fluency (TOWRE); WRMT = passage comprehension Woodcock Reading Mastery Test—Revised; RLN = rapid naming of letters; RWORD = real word recall; PWORD = pseudoword recall; DIG-F = digits forward task; DIG-B = digits backward task; LISSPAN = Listening Span task; SEMSET = Semantic Association Span task; AUDSET = auditory Digit/Sentence Span task; UPDATE = Updating Task; MATRIX = visual-spatial matrix task; MAP-DIR = mapping and direction task; CATF = categorical fluency; LETF = letter fluency; RANDL = random generation of letters; RANDN = random generation of numbers.

Received August 24, 2006

Revision received August 21, 2007

Accepted August 28, 2007 ■



# Revising the Redundancy Principle in Multimedia Learning

Richard E. Mayer and Cheryl I. Johnson  
University of California, Santa Barbara

College students viewed a short multimedia PowerPoint presentation consisting of 16 narrated slides explaining lightning formation (Experiment 1) or 8 narrated slides explaining how a car's braking system works (Experiment 2). Each slide appeared for approximately 8–10 s and contained a diagram along with 1–2 sentences of narration spoken in a female voice. For some students (the redundant group), each slide also contained 2–3 printed words that were identical to the words in the narration, conveyed the main event described in the narration, and were placed next to the corresponding portion of the diagram. For other students (the nonredundant group), no on-screen text was presented. Results showed that the group whose presentation included short redundant phrases within the diagram outperformed the nonredundant group on a subsequent test of retention ( $d = 0.47$  and  $0.70$ , respectively) but not on transfer. Results are explained by R. E. Mayer's (2001, 2005a) cognitive theory of multimedia learning, in which the redundant text served to guide the learner's attention without priming extraneous processing.

**Keywords:** educational technology, multimedia learning, redundancy effect, PowerPoint presentation

Suppose you want to explain how lightning storms develop or how a car's braking system works to students who lack relevant prior knowledge, so you design a concise, narrated animation using the following research-based principles of effective instructional design (Fletcher & Tobias, 2005; Mayer, 2001, 2005b, 2005c, 2005d):

1. The *multimedia principle*—you use both words (as spoken text) and pictures (as animation or a series of still frames).
2. The *coherence principle*—you minimize any extraneous words or pictures.
3. The *modality principle*—you present the words as narration rather than as on-screen text.
4. The *temporal contiguity principle*—you present the narration at the same time the corresponding event is depicted in the graphics.

On subsequent retention tests (e.g., in which you ask the learner to write an explanation) or transfer tests (e.g., in which you ask the learner to write why the system might not work or how to improve it), learners perform better than if the principles were not implemented (e.g., no pictures were presented, extraneous words were included in the narration, or the narration was presented before or after the graphics).

What can you do to improve upon such seemingly effective lessons? You might be tempted to incorporate on-screen text, in which a caption appears at the bottom of the screen that contains the same words that are being spoken in the narration. As soon as the narrator begins a sentence, it appears on the bottom of the

screen, and it stays on the screen until the narrator finishes the sentence.

Does adding redundant on-screen text help students learn? Previous research has shown students learn better from multimedia lessons containing graphics and narration than from graphics, narration, and redundant on-screen text (Kalyuga, Chandler, & Sweller, 1999, 2000, 2004; Leahy, Chandler, & Sweller, 2003; Mayer, Heiser, & Lonn, 2001; Moreno & Mayer, 2002a, 2002b; Mousavi, Low, & Sweller, 1995). This finding is known as the *redundancy effect* (Mayer, 2001, 2005c). For example, in a study by Moreno and Mayer (2002b), participants viewed an animation about lightning formation. The first condition had narration accompany the animation, whereas a second condition received redundant on-screen text in addition to the animation and narration. The group that received the redundant on-screen text performed worse on subsequent retention and transfer questions than did the group that received animation and narration; thus, a redundancy effect was found.

In most of the studies investigating the redundancy effect, including the one just discussed, the narration and the on-screen text were identical (Mayer, 2005c; Sweller, 1999, 2005), thus violating the coherence principle which, as mentioned above, states that unnecessary words or graphics should be eliminated (Mayer, 2001, 2005c). In addition, in many of the studies, including the one just discussed, the text was presented at the bottom of the screen, thus violating the *spatial contiguity principle*, which states that corresponding words and pictures in a multimedia presentation should be presented near each other on the screen (Ayres & Sweller, 2005; Mayer, 2001, 2005c).

Why does redundancy hinder learning? According to the cognitive theory of multimedia learning (Mayer, 2001, 2005a) shown in Figure 1, meaningful learning occurs when learners are able to pay attention to relevant portions of the words and graphics as they are registered in sensory memory (i.e., indicated by the "selecting words" and "selecting images" arrows), mentally organize them into coherent cognitive structures in working memory (i.e., indi-

---

Richard E. Mayer and Cheryl I. Johnson, Department of Psychology, University of California, Santa Barbara.

This project was supported by a grant from the Office of Naval Research.

Correspondence concerning this article should be addressed to Richard E. Mayer, Department of Psychology, University of California, Santa Barbara, Santa Barbara, CA 93106-9660. E-mail: mayer@psych.ucsb.edu

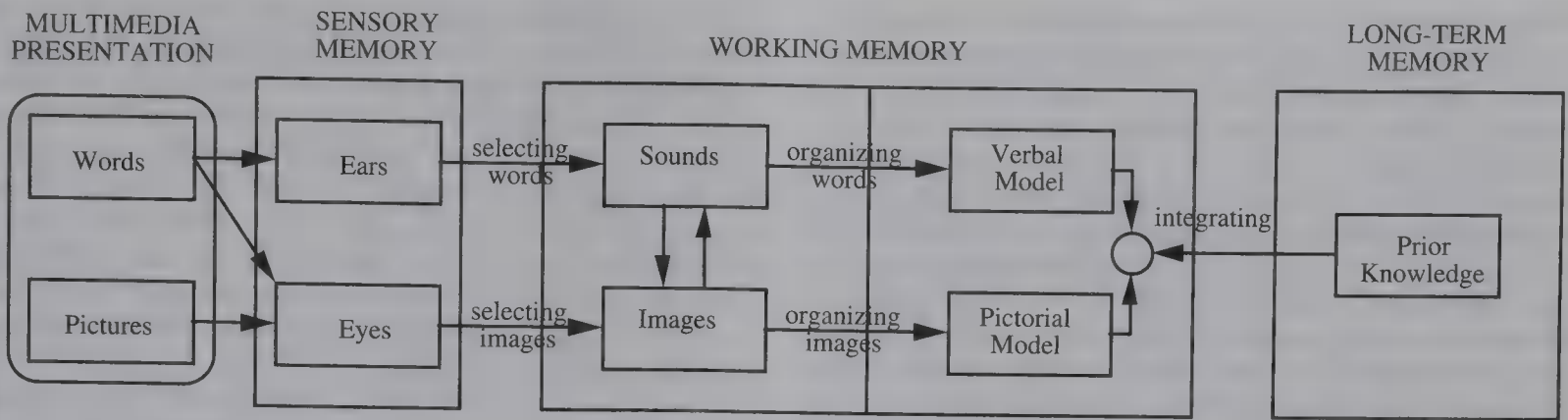


Figure 1. Cognitive theory of multimedia learning.

cated by the “organizing words” and “organizing images” arrows), and connect the verbal and pictorial representations with each other and with relevant knowledge retrieved from long-term memory (i.e., indicated by the “integrating” arrows). Although the integrating arrows are shown in working memory, the model assumes that the newly constructed knowledge is stored in long-term memory (using a process that can be called *encoding*).

Overall, the instructional message should be designed to reduce extraneous processing (i.e., processing that does not contribute to learning, such as visual scanning between the printed captions and the graphics), manage essential processing (i.e., processing aimed at selecting the relevant words and images so they can be represented in working memory), and foster generative processing (i.e., deeper processing in which the learner organizes and integrates the material). Adding redundant on-screen text detracts from the learning processes highlighted in Figure 1 because it creates extraneous processing—such as inducing the learner to visually scan between the caption at the bottom of the screen and the graphic and to try to mentally reconcile the incoming spoken and verbal stream. If the learner has to waste limited cognitive capacity on extraneous processing, the learner will be less able to engage in the cognitive processing needed for learning—essential and generative processing.

In spite of these empirical and theoretical setbacks for adding redundant on-screen text, in the current study we sought to determine whether there are conditions under which redundant on-screen text can foster rather than hinder learning. Consider a multimedia presentation on lightning formation consisting of 16 PowerPoint slides, each accompanied by approximately 10 s of narration, such as exemplified in the left panel of Figure 2 (i.e., the nonredundant presentation). To help learners direct their auditory attention to the key event being described in the narration, we can print a two- or three-word description of the event on the screen (using words from the narration). To help learners direct their visual attention to the key portion of the diagram, we can place the words next to the event they describe in the diagram. The right panel of Figure 2 exemplifies how we implemented the redundant words in the diagram (i.e., the redundant presentation). The goal of this manipulation is to facilitate the first step in the cognitive theory of multimedia learning—*selecting relevant words and images* (Mayer, 2001, 2005a). Success in selecting relevant words and images is most important for subsequent tests of retention of the presented material. Thus, based on the cognitive theory of multimedia learning, we predicted that the redundant group would outperform the nonredundant group on a retention test.

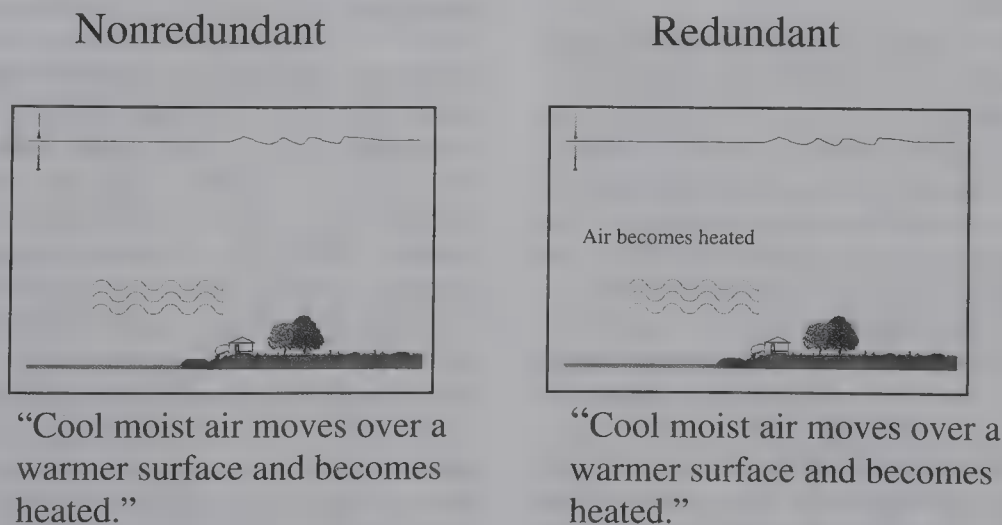


Figure 2. Example slide from the nonredundant (left panel) and redundant (right panel) versions of the lightning lesson.



This implementation of redundancy as short, on-screen labels avoids two major problems with previously studied redundancy presentations in which the entire text of the narration is printed at the bottom of the screen. First, by using only two or three keywords, the learner does not have to engage in reading the entire sentence; second, by placing the words next to the corresponding elements in the diagram, the learner does not need to scan from the bottom of the screen to various points in the diagram. Thus, we have attempted to create redundancy in a way that minimizes extraneous cognitive processing (namely, not requiring the learner to read an entire sentence or visually scan the diagram) while facilitating essential cognitive processing (namely, guiding the learner's selection of relevant words and images). Based on this interpretation of the cognitive theory of multimedia learning, we predict that adding on-screen labels to narrated graphics will result in improved performance on tests of retention (which are intended to be sensitive to instruction that guides the learner's attention) and will not hinder performance on tests of transfer (which are intended to measure deeper processing during learning). It is possible that the treatment could improve performance on transfer by freeing up cognitive capacity to engage in deeper processing, although the main manipulation in this study is intended to pinpoint the most important words in the narration.

### Experiment 1 (Lightning)

The purpose of Experiment 1 was to determine the cognitive consequences of adding short, redundant on-screen text to a multimedia lesson.

#### *Participants and Design*

Ninety undergraduates from the University of California, Santa Barbara, psychology student subject pool participated in this study for course credit. The mean age was 18.29 years ( $SD = .75$ ), and there were 58 women and 32 men. Forty-five participants served in the nonredundant group, and 45 participants served in the redundant group.

#### *Method*

**Materials and apparatus.** The paper-based materials consisted of a participant questionnaire, a retention test question, and four transfer test questions, each typed on an 8.5- × 11-in. sheet of paper. The participant questionnaire solicited information concerning the participant's age, sex, and prior knowledge. The retention test question was "Please write down an explanation of how lightning works." The retention test was intended to measure the student's memory for the presented material—corresponding to remembering factual and conceptual knowledge in Bloom's taxonomy (Anderson et al., 2001). The transfer test questions were as follows: (a) "What could you do to decrease the intensity of lightning?" (b) "Suppose you see clouds in the sky but no lightning. Why not?" (c) "What does air temperature have to do with lightning?" and (d) "What causes lightning?" The transfer test was intended to measure the student's understanding of the presented material—corresponding to understanding and applying conceptual knowledge in Bloom's taxonomy (Anderson et al., 2001).

The computer-based materials consisted of nonredundant and redundant versions of a PowerPoint lesson on lightning formation. The presentation included diagrams and simultaneous narration shown via PowerPoint and consisted of 16 individual slides, lasting approximately 2.5 min. The slides advanced automatically after the narration was complete, averaging approximately 9.5 s per slide. The slides for the redundant group included two or three action-oriented keywords from the narration and were presented within the diagram next to the elements they described. The words expressed the main action depicted in the slide. For example, for the slide containing the narration, "Cool moist air moves over a warmer surface and becomes heated," the keywords "air becomes heated" were printed in the diagram next to the heated air mass. The slides for the nonredundant group contained no printed words. Examples of redundant and nonredundant slides are shown in Figure 2.

The apparatus consisted of five Macintosh iBook computers with Panasonic headphones. A stopwatch was used to time the retention and transfer tests.

**Procedure.** Students were tested in groups of 1 to 5 per session, with each student seated in an individual cubicle containing a Macintosh iBook computer with Panasonic headphones. Students were told they would view a lesson on lightning formation and then they would answer some questions based on the material. Each student completed the participant questionnaire, then watched a short PowerPoint presentation about lightning formation on the computer, and finally answered a series of timed, open-ended questions about the material consisting of one retention question and four transfer questions. Half of the participants were randomly assigned to the nonredundant group (which received graphics and narration), and half were randomly assigned to the redundant group (which received graphics, narration, and on-screen text). Following the presentation, all students wrote their answers for the retention test question (with a 5-min time limit) and each of the transfer test questions (with a 3-min time limit per question). Upon completion of the tests, students were thanked and debriefed. The materials and procedure were adapted from previous research using a narrated animation on lightning formation (Mayer et al., 2001; Moreno & Mayer, 2002b).

#### *Results and Discussion*

The dependent variables were the scores on the retention and transfer tests; the independent variable was whether or not the presentation included redundant text. In order to score the retention test, we broke the lightning script into 16 idea units, each expressing a main event such as "cool air moves over warmer surface and becomes heated." Each student's retention test was scored by assigning 1 point for each idea unit that the student wrote down, regardless of wording or spelling (out of a total possible of 16 points), such as "cold air gets hotter." For the transfer tests, a list of acceptable answers was created for each question. For example, for the question about decreasing the intensity of lightning, correct answers included removing positive ions from the ground and placing positive ions near the cloud. One point was awarded for each correct answer, and the points for each question were added up to compute the total transfer score (out of a total possible of 12 points), which was used in the analyses.

Table 1 presents the means and standard deviations of the two groups on the retention and transfer tests. Independent samples *t* tests revealed that the redundant group ( $M = 8.02$ ,  $SD = 2.83$ ) scored significantly better than the nonredundant group ( $M = 6.69$ ,  $SD = 2.87$ ) on the retention test,  $t(88) = 2.22$ ,  $p < .05$ ,  $d = 0.47$ , whereas the redundant group ( $M = 3.60$ ,  $SD = 2.00$ ) and the nonredundant group ( $M = 3.67$ ,  $SD = 2.04$ ) did not differ significantly on the transfer test,  $t(88) = 0.16$ ,  $p = .88$ ,  $d = 0.04$ .

Results can be explained in terms of Mayer's (2001, 2005a) cognitive theory of multimedia learning. Essential processing was fostered by the redundant treatment because the on-screen text helped guide the learners' attention to the relevant words—which was reflected in retention test performance. However, the redundant on-screen text was not intended to encourage generative processing (i.e., deeper cognitive processing), such as organizing the material into a coherent representation and integrating it with existing knowledge—which would be reflected in transfer test performance. Extraneous processing was kept low in the redundant condition by minimizing the amount of on-screen text (rather than reproducing all of the text) and placing it next to the corresponding portion of the diagram (rather than at the bottom).

### Experiment 2 (Brakes)

The purpose of Experiment 2 was to determine whether the pattern of results obtained in Experiment 1 could be replicated with different instructional materials.

#### Participants and Design

Sixty-two undergraduate students from the University of California, Santa Barbara, psychology student subject pool participated in this study for course credit. The mean age of participants was 18.6 years ( $SD = 1.2$ ), and there were 27 women and 35 men. Thirty-one participants served in the nonredundant group, and 31 participants served in the redundant group.

#### Method

**Materials and apparatus.** The paper-based materials consisted of a participant questionnaire, a retention test question, and five transfer test questions, each typed on an 8.5- × 11-in. sheet of paper. The participant questionnaire solicited information concerning the participant's age, sex, and prior knowledge. The retention test question was "Please write down an explanation for how a car's braking system works." The transfer test questions were as

follows: (a) "Why do brakes get hot?" (b) "What could be done to make brakes more reliable, that is, to make sure they do not fail?" (c) "What could be done to make brakes more effective, that is, to reduce the distance needed to bring the car to a stop?" (d) "Suppose you press on the brake pedal in your car but the brakes don't work. What could have gone wrong?" and (e) "What happens when you pump the brakes (i.e., press the pedal and release the pedal repeatedly and rapidly)?"

The computer-based materials consisted of nonredundant and redundant versions of a PowerPoint lesson on how a car's braking system works. The presentation included diagrams and simultaneous narration shown via PowerPoint and consisted of eight individual slides, lasting approximately 80 s. The slides advanced automatically after the narration was complete, averaging approximately 10 s per slide. The slides for the redundant group included two or three action-oriented keywords from the narration and were presented within the diagram next to the elements they described. The words expressed the main action in the slide. For example, for the slide containing the narration, "a piston moves forward inside the master cylinder," the keywords "piston moves forward" were printed in the diagram next to the piston in the master cylinder. The slides for the nonredundant group contained no printed words. An example of a redundant slide is presented in Figure 3.

The apparatus consisted of five Macintosh iBook computers with Panasonic headphones. A stopwatch was used to time the retention and transfer tests.

**Procedure.** Students were tested in groups of 1 to 5 per session, with each student seated in an individual cubicle containing a Macintosh iBook computer with Panasonic headphones. Students were told they would view a lesson on how car brakes work, and then would answer some questions based on the material in the lesson. Each student completed the participant questionnaire, then watched a short PowerPoint presentation about car brakes on the computer, and finally answered a series of timed, open-ended questions about the material consisting of one retention question and five transfer questions. Half of the participants were randomly assigned to the nonredundant group (which received graphics and narration), and half were randomly assigned to the redundant group (which received graphics, narration, and on-screen text). Following the presentation, all students wrote their answers for the retention test question (with a 4-min time limit) and each of the transfer test questions (with a 2.5-min time limit per question). Upon completion of the tests, students were thanked and debriefed. The materials and procedure were adapted from previous research using a narrated animation on how brakes work (Mayer & Anderson, 1991; Mayer & Moreno, 1998).

#### Results and Discussion

The dependent variables were the scores on the retention and transfer tests; the independent variable was whether or not the presentation included redundant text. The retention tests were scored by breaking the lesson into eight idea units, each conveying one action, such as "the piston moves forward in the master cylinder." Students received 1 point for each idea unit they wrote down on their retention test (out of a total possible of 8 points), regardless of spelling or wording. For the transfer tests, we developed a list of acceptable answers for each question. For example, for the question about what could be wrong, possible answers

Table 1  
Mean Retention and Transfer Scores for Two  
Groups—Experiment 1 (Lightning Lesson)

Score type and group	<i>M</i>	<i>SD</i>	<i>p</i>	<i>d</i>
Retention				
Nonredundant	6.7	2.9	.03	0.47
Redundant	8.0	2.8		
Transfer				
Nonredundant	3.6	2.0	.88	0.04
Redundant	3.7	2.0		



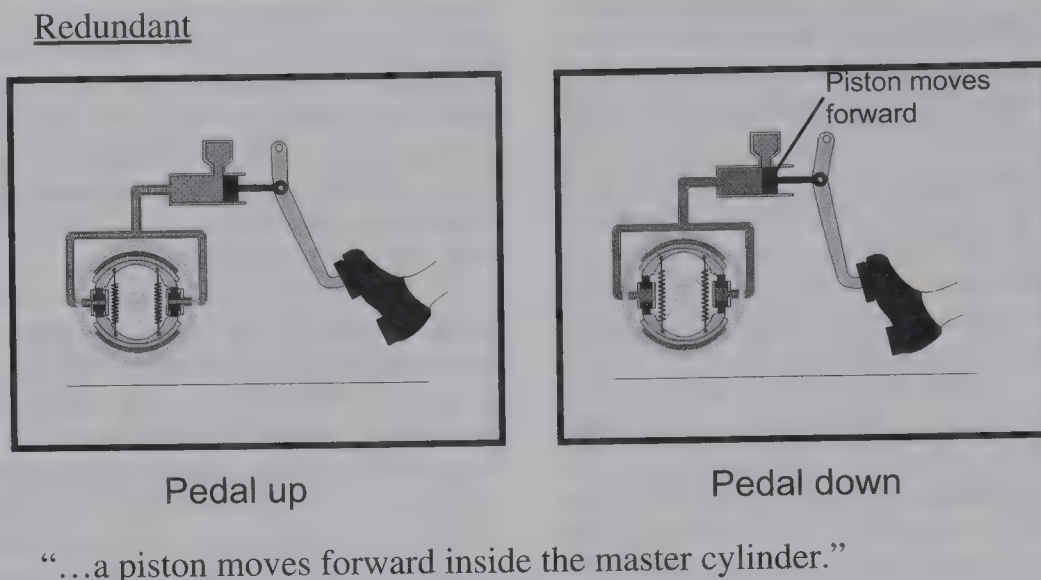


Figure 3. Slide from the redundant version of the brake lesson; the nonredundant version did not contain the on-screen text.

included that the piston was stuck or there was a leak in the brake tube. One point was awarded for each correct answer, and the points for each question were added up to compute the total transfer score (out of 14 possible points), which was used in the analyses.

Table 2 summarizes the means and standard deviations for the two groups on the retention and transfer tests. Independent samples *t* tests revealed that the redundant group ( $M = 5.13$ ,  $SD = 1.98$ ) scored significantly better than the nonredundant group ( $M = 3.61$ ,  $SD = 2.35$ ) on the retention test,  $t(60) = 2.75$ ,  $p = .008$ ,  $d = 0.70$ , whereas the redundant group ( $M = 3.90$ ,  $SD = 1.76$ ) and the nonredundant group ( $M = 4.16$ ,  $SD = 2.16$ ) did not differ significantly on the transfer test,  $t(60) = 0.52$ ,  $p = .608$ ,  $d = 0.15$ .

The pattern of results in Experiment 2 replicates that in Experiment 1 and supports the theoretical explanation that short, redundant on-screen text can guide the learner's attention toward relevant words (i.e., essential processing) without increasing extraneous processing or fostering generative processing.

## Conclusion

### Main Empirical Findings

In two separate experiments, adding short, redundant text to narrated graphics resulted in improvements on retention of the

verbal material but not on transfer. The effect size—favoring the redundant group—was in the medium range, indicating that the redundancy effect was practically important as well as statistically significant.

### Theoretical Implications

The pattern of results is consistent with the predictions of the cognitive theory of multimedia learning, in which short on-screen labels are intended to guide the cognitive process of selecting relevant words and images while not creating extraneous processing. In previous studies, adding redundant on-screen text to narrated graphics (i.e., complete reproductions of the narration as captions at the bottom of the screen) resulted in decreases in both retention and transfer test performance (Mayer, 2001, 2005c).

There were three main differences in the way redundancy was implemented in the present study in comparison with previous studies that yielded a redundancy effect. First, the on-screen text was short rather than long: Short text, consisting of two or three words, was intended to guide essential processing while not creating extraneous processing, whereas long on-screen text does not guide essential processing while creating the need for extraneous processing.

Second, the on-screen text was placed near rather than far from the corresponding portion of the graphic: Contiguous presentation of corresponding words and graphics is intended to minimize extraneous processing while guiding the learner's attention toward the relevant portion of the graphic, whereas separated presentation induces extraneous processing while not guiding the learner's attention. When sentences from the narration are printed in their entirety as text at the bottom of the screen, learners may engage in extraneous cognitive processing that results in poorer performance on subsequent tests (Mayer, 2001, 2005c). In contrast, when a few keywords are printed next to the corresponding portion of the diagram, learners may be better able to engage in essential cognitive processing (such as selecting relevant words and images) that results in improved performance on subsequent tests.

Table 2  
Mean Retention and Transfer Scores for Two Groups—Experiment 2 (Brakes Lesson)

Score type and group	<i>M</i>	<i>SD</i>	<i>p</i>	<i>d</i>
Retention				
Nonredundant	3.6	2.3	.008	0.70
Redundant	5.1	2.0		
Transfer				
Nonredundant	4.2	2.2	.61	0.15
Redundant	3.9	1.8		

Table 3  
When Redundancy Supports or Hinders Learning

Action	Condition
Redundancy supports learning	When it minimizes extraneous processing—by placing text near rather than far from corresponding graphics When it fosters essential processing—by highlighting key portions of the text rather than presenting all text (or when text is complex)
Redundancy hinders learning	When it creates extraneous processing—by placing text far from rather than near corresponding graphics When it detracts from essential processing—by highlighting the entire text rather than highlighting key portions of the text

Third, the present study presented graphics as a series of static illustrations rather than as an animation as in previous experiments. The motion in an animation may help direct the learner's attention to the relevant portion of the graphic, but there are no motion cues in a static illustration. Therefore, redundant on-screen phrases may be particularly helpful for directing the learner's attention with static illustrations.

### Practical Implications

Concerning educational implications, this study helps to establish some boundary conditions for determining when redundancy can be helpful or harmful in multimedia learning. The most straightforward contribution of this work is to add an important caveat to the redundancy principle "People learn more deeply from graphics and narration than from graphics, narration, and on-screen text" (Mayer, 2005c, p. 193).<sup>1</sup> In revising this statement of the redundancy principle, we can add the following limitation: "except when the on-screen text is short, highlights the key action described in the narration, and is placed next to the portion of the graphic that it describes." In short, this work pinpoints an important boundary condition for applying the redundancy principle—it applies most strongly when the on-screen text reproduces the complete narration and appears at the bottom of the screen, and it does not necessarily apply when the on-screen text consists of a few core words placed next to the relevant portion of the graphic. Table 3 summarizes the conditions under which redundancy can hinder or help learning. We also suspect that other boundary conditions include when the spoken text is complex, contains unfamiliar words, or is not in the learner's native language; when the pace of the presentation is slow or under the learner's control; or when no graphics are presented (Mayer, 2001, 2005c).

A broader contribution of this work is to highlight the admonition that design principles for multimedia messages, such as the 10 principles proposed by Mayer (2001, 2005b, 2005c, 2000d), are not rigid laws that must be followed in all circumstances. Rather, decisions about appropriate instructional design should be based on an understanding of how people learn from words and pictures, such as the cognitive theory of multimedia learning (Mayer, 2001, 2005a) or cognitive load theory (Sweller, 1999, 2005). Thus, the challenge for instructional designers is to apply design principles in ways that reduce extraneous processing (such as scanning between captions and the graphic), manage intrinsic processing (such as attending to relevant portions of the narration and graphic), and foster generative processing (such as mentally organizing and integrating the material). The revision to the redundancy principle is necessary because the use of short labels helps

manage essential processing (by guiding the learner's attention to the keywords in the narration and the key action in the graphic) while not adding to extraneous processing (by presenting only a few words and placing them next to the portion of graphic they describe). In summary, rather than blindly following design rules, instructional designers should always consider how applying a rule will affect the learner's cognitive processing during learning, particularly the degree to which applying the principle is likely to lead to reducing extraneous processing, managing intrinsic processing, and fostering generative processing.

### Limitations and Future Directions

These studies were conducted as short laboratory experiments, but future work is needed to determine whether the findings generalize to more realistic educational settings. Future work also is needed to determine whether the results would be obtained if the words in the narration were presented as on-screen captions rather than narration. Finally, although there were no differences between the groups on the transfer test, it is possible that more sensitive measures might be able to detect a difference in future work. However, the transfer scores were not near either ceiling or floor, and the same transfer test has yielded significant differences between treatment groups in other experiments (Mayer, 2001).

<sup>1</sup> Sweller (1999, 2005) uses a broader definition of the redundancy principle in which learning is hindered when "the same material is presented in multiple forms" (Sweller, 2005, p. 159). This definition includes our definition of redundancy but also includes additional situations that we did not investigate.

### References

- Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., et al. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.
- Ayres, P., & Sweller, J. (2005). The split attention effect in multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 135–146). New York: Cambridge University Press.
- Fletcher, J. D., & Tobias, S. (2005). The multimedia principle. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 117–134). New York: Cambridge University Press.
- Kalyuga, S., Chandler, P., & Sweller, J. (2004). When redundant on-screen text in multimedia technical instruction can interfere with learning. *Human Factors*, 46, 567–581.
- Kalyuga, S., Chandler, P., & Sweller, P. (1999). Managing split-attention



- and redundancy in multimedia instruction. *Applied Cognitive Psychology*, 13, 351–372.
- Kalyuga, S., Chandler, P., & Sweller, P. (2000). Incorporating learner experience into the design of multimedia instruction. *Journal of Educational Psychology*, 92, 126–136.
- Leahy, W., Chandler, P., & Sweller, P. (2003). When auditory presentation should and should not be a component of multimedia instruction. *Applied Cognitive Psychology*, 17, 401–418.
- Mayer, R. E. (2001). *Multimedia learning*. New York: Cambridge University Press.
- Mayer, R. E. (2005a). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 31–48). New York: Cambridge University Press.
- Mayer, R. E. (2005b). Principles for managing essential processing in multimedia learning: Segmenting, pretraining, and modality. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 169–182). New York: Cambridge University Press.
- Mayer, R. E. (2005c). Principles for reducing extraneous processing in multimedia learning: Coherence, signaling, redundancy, spatial contiguity, and temporal contiguity. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 183–200). New York: Cambridge University Press.
- Mayer, R. E. (2005d). Principles of multimedia learning based on social cues: Personalization, voice, and image principles. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 201–212). New York: Cambridge University Press.
- Mayer, R. E., & Anderson, R. B. (1991). Animations need narrations: An experimental test of a dual-coding hypothesis. *Journal of Educational Psychology*, 83, 484–490.
- Mayer, R. E., Heiser, H., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology*, 93, 187–198.
- Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. *Journal of Educational Psychology*, 90, 312–320.
- Moreno, R., & Mayer, R. E. (2002a). Learning science in virtual reality multimedia environments: Role of method and media. *Journal of Educational Psychology*, 94, 598–610.
- Moreno, R., & Mayer, R. E. (2002b). Verbal redundancy in multimedia learning: When reading helps listening. *Journal of Educational Psychology*, 94, 156–163.
- Mousavi, S. Y., Low, R., & Sweller, J. (1995). Reducing cognitive load by mixing auditory and visual presentation modes. *Journal of Educational Psychology*, 87, 319–334.
- Sweller, J. (1999). *Instructional design in technical areas*. Camberwell, Australia: ACER Press.
- Sweller, J. (2005). The redundancy principle in multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 159–168). New York: Cambridge University Press.

Received May 8, 2007

Revision received August 15, 2007

Accepted October 9, 2007 ■

## Low Publication Prices for APA Members and Affiliates

**Keeping you up-to-date.** All APA Fellows, Members, Associates, and Student Affiliates receive—as part of their annual dues—subscriptions to the *American Psychologist* and *APA Monitor*. High School Teacher and International Affiliates receive subscriptions to the *APA Monitor*, and they may subscribe to the *American Psychologist* at a significantly reduced rate. In addition, all Members and Student Affiliates are eligible for savings of up to 60% (plus a journal credit) on all other APA journals, as well as significant discounts on subscriptions from cooperating societies and publishers (e.g., the American Association for Counseling and Development, Academic Press, and Human Sciences Press).

**Essential resources.** APA members and affiliates receive special rates for purchases of APA books, including the *Publication Manual of the American Psychological Association*, and on dozens of new topical books each year.

**Other benefits of membership.** Membership in APA also provides eligibility for competitive insurance plans, continuing education programs, reduced APA convention fees, and specialty divisions.

**More information.** Write to American Psychological Association, Membership Services, 750 First Street, NE, Washington, DC 20002-4242.

# Does Extrinsic Goal Framing Enhance Extrinsic Goal-Oriented Individuals' Learning and Performance? An Experimental Test of the Match Perspective Versus Self-Determination Theory

Maarten Vansteenkiste  
Ghent University

Tinneke Timmermans and Willy Lens  
University of Leuven

Bart Soenens  
Ghent University

Anja Van den Broeck  
University of Leuven

Previous work within self-determination theory has shown that experimentally framing a learning activity in terms of extrinsic rather than intrinsic goals results in poorer conceptual learning and performance, presumably because extrinsic goal framing detracts attention from the learning activity and is less directly satisfying of basic psychological needs. According to the match perspective, experimental extrinsic, compared to intrinsic, goal framing should enhance learning and performance for learners who personally hold a stronger extrinsic than intrinsic goal orientation, as these learners' personally held goals match with the situationally induced goals. An experimental field study among 5th–6th grade children shows that extrinsic goal framing resulted in poorer autonomous motivation, conceptual (but not rote) learning, and persistence compared to intrinsic goal framing, irrespective of participants' personal intrinsic versus extrinsic goal orientations and their spontaneous perception of the learning activity as serving an intrinsic or an extrinsic goal. The authors conclude that teachers can best promote intrinsic goals, even when facing students who personally hold a stronger extrinsic than intrinsic goal orientation.

**Keywords:** intrinsic versus extrinsic goals, match perspective, self-determination theory

Values represent categories of desirable life goals, varying in importance, that serve as a guiding principle in people's lives (De Witte, 2004; Rohan, 2000; Rokeach, 1973; Schwartz, 1992, 1999). Various researchers have developed multidimensional models of the interpersonal variation in the type of values (Schwartz, 1994) or life goals (Grouzet et al., 2006; Nuttin & Lens, 1985) that people hold.<sup>1</sup> In addition, researchers have been interested in studying the implications of people's life goals for their well-being and achievement. In doing so, they have proposed at least two different and conflicting perspectives. Some researchers, drawing on SDT (Deci & Ryan, 1985, 2000; Ryan & Deci, 2000b), hold that some goals are better than others for people's thriving and optimal functioning (Kasser, 2002; Vansteenkiste, Lens, & Deci, 2006). Other researchers claim that there does not exist a category of more adaptive values or life goals (e.g., Bianco, Higgins, & Klem, 2003; Sagiv & Schwartz, 2000; Walsh & Holland, 1992). They have argued that the relation between people's life goals,

achievement, and adjustment depends on the match or fit between people's personal goal orientations and the goals that are emphasized and encouraged in the direct environment (Pervin, 1968; Schneider, 1987). This viewpoint is commonly referred to as the *match perspective*.

A few previous studies have directly examined the conflicting hypotheses that can be derived from SDT and the match perspective (Kasser & Ahuvia, 2002; Sagiv & Schwartz, 2000; Vansteenkiste, Duriez, Simons, & Soenens, 2006), but they were all correlational in nature and focused on well-being outcomes only. The current contribution goes beyond past work by providing evidence from an experimental field study that focused on learning and performance outcomes. Before detailing the aims of the present research, we introduce both theoretical perspectives.

---

Maarten Vansteenkiste and Bart Soenens, Department of Psychology, Ghent University, Ghent, Belgium; Tinneke Timmermans, Willy Lens, and Anja Van den Broeck, Department of Psychology, University of Leuven, Leuven, Belgium.

Anja Van den Broeck's contributions were supported by the Grant for Scientific Research Flanders.

Correspondence concerning this article should be addressed to Maarten Vansteenkiste, H. Dunantlaan 2, 9000, Ghent, Belgium. E-mail: maarten.vansteenkiste@ugent.be

---

<sup>1</sup> It should be noted that there exists a slight conceptual difference between values and life goals, as the latter are more concrete and lower order units, whereas the former are rather abstract and higher order units (Grouzet et al., 2006; Schwartz, 1994). The conceptual difference between the constructs is, however, of less importance to the present study, which focuses on the match between personally held life goals or values and the type of life goals or values that are promoted by the social environment. We adopted the term *goals* throughout the article because this is the term that is most commonly used within research on self-determination theory (SDT).



## Self-Determination Theory

### *Intrinsic and Extrinsic Motivation and Intrinsic and Extrinsic Goal Contents*

The classic motivation constructs that have received the most empirical and theoretical attention within SDT and other motivational frameworks are intrinsic and extrinsic motivation (Deci & Ryan, 2000; Vansteenkiste, Lens, & Deci, 2006). Whereas *intrinsic motivation* refers to engaging in an activity for its own sake, simply because it is enjoyable and gratifying by itself, *extrinsic motivation* refers to engaging in an activity to obtain an outcome separable from the activity itself (Ryan & Deci, 2000a). In earlier theorizing in the 1970s (e.g., Lepper & Greene, 1978), intrinsic and extrinsic motivation were considered as antipodes, with intrinsic motivation representing the most autonomous or self-determined type of motivation and extrinsic motivation representing a nonautonomous or controlled type of motivation. However, subsequent empirical work in the 1980s (e.g., Ryan, Mims, & Koestner, 1983) made clear that not all extrinsically motivated behavior is nonautonomous, given that individuals are able to volitionally endorse (i.e., internalize) the reasons for engaging in the activity at hand. This insight resulted in a conceptual shift within SDT away from the intrinsic–extrinsic motivation dichotomy toward the distinction between autonomous motivation, which involves both intrinsic and well-internalized motivation, and controlled motivation, which involves poorly internalized or non-internalized motivation. Thus, autonomous and controlled motivation represent two qualitatively different types of reasons for regulating one's behavior and are often studied under the label of the “why” of behavior (Deci & Ryan, 2000; Vansteenkiste, Lens, & Deci, 2006; Vansteenkiste, Ryan, & Deci, in press).

Parallel to the growing attention to the concepts of autonomous and controlled motivation, SDT researchers have also been paying attention to people's goal contents or goal orientations (Kasser & Ryan, 1993). In this respect, Kasser and Ryan (1993, 1996) differentiated extrinsic goals, such as wealth, physical appeal, and social recognition, from intrinsic goals, such as affiliation, personal growth, and community contribution, on the basis of humanistic (Maslow, 1954) and existential (Fromm, 1976) theorizing. Intrinsic goals reflect people's natural growth tendencies and are characterized by an inwardly oriented frame. In contrast, when people pursue extrinsic goals, they tend to adopt an outward orientation (Williams, Cox, Hedberg, & Deci, 2000)—that is, they are focused on impressing others by garnering external signs of self-worth. These different goal contents are often studied under the label of the “what” of behavior within SDT (Deci & Ryan, 2000).

According to SDT, intrinsic and extrinsic goals, which represent qualitatively different types of goals, can be motivated by autonomous or controlled reasons, which represent qualitatively different types of motivations and thus distinct sets of constructs. For instance, a person can be altruistic (i.e., an intrinsic goal) because he or she enjoys pursuing this goal (autonomous motivation) or because he or she would feel bad if he or she did not help people in need (controlled motivation). Similarly, a person can be focused on becoming famous (i.e., an extrinsic goal) because his or her parents pressure him or her to be a well-known person (controlled motivation) or because he or she personally values fame (autonomous motivation).

In the present research, we explicitly focused on the concepts of intrinsic and extrinsic goal orientations. More specifically, we examined whether framing a learning activity in terms of the attainment of intrinsic or extrinsic goals would yield a differential impact on the learning and persistence of intrinsic and extrinsic goal-oriented individuals. To clarify this overall aim, we first elaborate more on the concepts of intrinsic and extrinsic goal contents and their effects, as they have been studied from the SDT perspective.

### *The Effects of Intrinsic Versus Extrinsic Goal Contents*

SDT starts with the central assumption that all individuals are born with the basic psychological needs for autonomy, competence, and relatedness. The satisfaction of these psychological needs is regarded as necessary nutrients for individuals' optimal performance, well-being, and development, much as plants need water and sunshine to flower (Deci & Ryan, 2000; Ryan & Deci, 2000b). To understand the effects of goal contents on people's optimal functioning, one needs to answer the question of how goal pursuit is related to the satisfaction of the basic psychological needs. In this respect, intrinsic goals are said to facilitate psychological well-being, because they provide direct satisfaction of the basic and universal psychological needs for competence, autonomy, and relatedness (Deci & Ryan, 2000; Kasser, 2002). For instance, when people endorse the intrinsic goal of self-development, they are more likely to engage in challenging activities in a task-focused manner, which increases their chances of gaining a sense of competence. When people value the development of close relationships, they are more likely to empathically take the perspective of others, which might facilitate a sense of mutual connectedness and thrust (Kasser, 2002; Vansteenkiste, Neyrinck, et al., 2007; Vansteenkiste, Soenens, & Lens, 2007).

Although extrinsic goal pursuit might provide some satisfaction, this type of satisfaction is likely to be derivative and short lived, because extrinsic goal pursuits do not directly satisfy individuals' basic psychological needs (Deci & Ryan, 2000; Kasser, 2002; Kasser, Ryan, Couchman, & Sheldon, 2004). For instance, extrinsic goal-oriented individuals tend to approach others in an “objectifying” manner (Kasser, 2002)—that is, to treat them as objects that should be used in the most efficient way toward one's extrinsic goal attainment. Further, (learning) activities are approached in a restricted and rigid manner (Vansteenkiste, Simons, Lens, Soenens, & Matos, 2005); that is, extrinsic goal-oriented individuals only put effort into the activity as far as this will help them in achieving their extrinsic ambitions. Thus, they are less deeply involved in learning tasks because engagement in such tasks is only valued to the extent that they are instrumental for reaching extrinsic goals (Vansteenkiste, Simons, Lens, Soenens, et al., 2004). Such an objectifying and narrow-focused approach is, however, unlikely to create opportunities for satisfaction of the basic psychological needs for relatedness, competence, and autonomy. Consistent with these claims, Vansteenkiste, Neyrinck, et al. (2007) found in a sample of employees that an extrinsic, relative to an intrinsic, goal pursuit hindered basic need satisfaction.

Because of their differential linkage to the satisfaction of basic psychological needs, intrinsic versus extrinsic goals differentially predict psychological well-being (Ryan, Sheldon, Kasser, & Deci, 1996). In line with this reasoning, previous research has docu-



mented that when participants' life aspirations are more strongly centered around the pursuit of extrinsic rather than intrinsic goals, they display lower psychological well-being and social adjustment and are more vulnerable to ill-being and drug and alcohol abuse (see Kasser, 2002; Kasser et al., 2004, for an overview). Although most of this work has been conducted among late adolescent and adult samples, a few studies have suggested that early adolescents (Cohen & Cohen, 1996) and children (Kasser, 2005) also suffer emotionally from adopting an extrinsic or materialistic goal orientation.

It is interesting that SDT maintains that the pursuit of extrinsic, relative to intrinsic, goals should hinder the learning process and individuals' broader psychosocial adjustment, even when extrinsic goals are highly emphasized by the social environment—that is, in spite of a person–environment match in extrinsic goals. This is because extrinsic goals, by their very nature, are less likely to facilitate the satisfaction of basic human (organismic) needs, which functions as the “input” for all individuals' optimal functioning. Consistent with this, Kasser and Ahuvia (2002) and Vansteenkiste, Duriez, et al. (2006) found that extrinsic, relative to intrinsic, goal pursuit was associated with lower well-being and more internal distress among a group of business students (see also Srivastava, Locke, & Bartol, 2001), even though extrinsic goals tend to be heavily emphasized in business environments (Holland, 1985).

More recently, Vansteenkiste and colleagues (see Vansteenkiste, Lens, & Deci, 2006, for an overview) extended this work by examining the impact on participants' learning and performance of experimentally framing a learning activity in terms of an intrinsic versus extrinsic goal attainment. Thus, instead of treating intrinsic versus extrinsic goal contents as an individual-differences characteristic, they studied goal contents from a social contextual viewpoint—that is, by considering the extent to which goal contents are promoted by the social environment. On the basis of SDT, it was reasoned that intrinsic goals concern learning in the service of inherent psychological needs and growth tendencies and, hence, should promote learning. In contrast, extrinsic goals, with their focus on external indicators of worth, are more likely to distract people from the learning activity and, hence, should result in poorer learning. In a series of experimental field studies, it was found that framing a learning activity in terms of the attainment of an intrinsic goal enhanced conceptual learning, performance, and persistence compared to (a) an extrinsic goal-framing condition (Vansteenkiste, Simons, Lens, Sheldon, & Deci, 2004), (b) a double (i.e., intrinsic plus extrinsic) goal-framing condition (Vansteenkiste, Simons, Lens, Soenens, et al., 2004), and (c) a no-goal control condition (Vansteenkiste, Simons, Soenens, & Lens, 2004), whereas extrinsic goal framing undermined learning and performance both compared to the double goal-framing condition and the no-goal control condition. These findings were replicated among 11–12-year-old children (Vansteenkiste et al., 2005).

Although research has shown that experimental extrinsic, relative to intrinsic, goal framing undermines conceptual learning and performance, the question should be raised regarding whether these negative effects also apply to individuals who are, by themselves, more strongly oriented toward the pursuit of extrinsic than intrinsic goals. On the basis of the match perspective, it can be suggested that the negative effects of extrinsic goal framing are limited to learners who value intrinsic goals over extrinsic goals

and that the overall enhancement of learning and persistence in the intrinsic goal conditions of the studies reviewed earlier was carried primarily by those learners whose goal orientation was more intrinsic than extrinsic. To better understand the hypothesis derived from the match perspective, we describe this perspective in more detail below.

### The Match Perspective

The match perspective posits that the effect of the content of people's goal pursuits depends on the match or fit between the goal content and the type of goals that are emphasized in the interpersonal environment (Sagiv & Schwartz, 2000). The theme of match or fit between the person and the environment is represented in a broad array of fields in psychological research, as diverse as social (e.g., Sagiv & Schwartz, 2000), educational (e.g., Harackiewicz & Elliot, 1998), organizational (e.g., Edwards, 1991; Meglino, Ravlin, & Adkins, 1989; Ostroff, 1993; Ostroff, Shin, & Kinicki, 2005; Pervin, 1968), developmental (e.g., Eccles, Lord, & Midgley, 1991), and sport (e.g., Amiot, Vallerand, & Blanchard, 2006) psychology. However, most of this work, except for Sagiv and Schwartz (2000), concerned the examination of a person–environment match in goals other than the intrinsic–extrinsic goal dimension outlined within SDT. Hence, the present research extends previous work on the match perspective by examining whether the match perspective also holds with regard to the theoretical distinction between intrinsic and extrinsic goals, which has received increasing empirical attention in the educational and motivational literature (see Vansteenkiste, Lens, & Deci, 2006).

Sagiv and Schwartz (2000) claimed that the negative effect of extrinsic, relative to intrinsic, goal pursuits should be reversed when people find themselves in an extrinsic goal context that emphasizes extrinsic goals over intrinsic goals. They provided three related mechanisms to explain this matching effect. First, social environments that match individuals' goals afford better opportunities to translate one's goals into plans, to carry out one's plans, and, hence, to attain one's goals. Second, when people reject the goals and values that prevail in their environment, they may be ignored, ostracized, or sanctioned, while acceptance of the socially promoted goals is likely to elicit social support and reinforcement, which should promote well-being. Third, the internal conflict that people experience when their earlier acquired goals are in conflict with the goals supported by the social environment is also likely to forestall their well-being.

In line with the match perspective, Sagiv and Schwartz (2000), relying on Schwartz's (1992) value classification, showed in a sample of psychology and business students that the effect of valuing extrinsic goals over intrinsic goals on well-being interacted with participants' study environment. Business students who valued extrinsic goals over intrinsic goals reported higher psychological well-being, while psychology students reported more optimal functioning when they valued intrinsic goals over extrinsic goals.

The present study builds on the work of Sagiv and Schwartz (2000), because it examined whether the experimental induction of an intrinsic versus an extrinsic goal prior to engagement in a learning activity differentially affected learning and performance among children who valued intrinsic goals over extrinsic goals versus children who valued extrinsic goals over intrinsic goals.



Although the three mechanisms reviewed above are primarily relevant for the issue of well-being, the third mechanism (i.e., internal conflict) might also help us to understand the effects of match versus mismatch on learning and performance. That is, the third mechanism suggests that when learners are placed in a mismatch condition, they face an internal conflict because their own personal goal orientation does not fit with the goal that is experimentally induced. Two groups of learners are likely to face a mismatch and, hence, to experience internal conflict: (a) learners who personally value extrinsic goals over intrinsic goals and are placed in an intrinsic goal condition, and (b) learners who personally value intrinsic goals over extrinsic goals and are placed in an extrinsic goal condition. The experience of internal conflict might result in a less concentrated and less task-involved approach to the learning activity and, hence, help to explain why the mismatch, versus the match, conditions should undermine performance and persistence.

In addition to examining whether a match between one's personal goal orientation and the goal-framing condition yields more optimal learning, we also examined whether being placed in an experimental goal-framing condition that matches with one's spontaneous perception of the learning activity promotes optimal learning. Specifically, herein we suggest that there exists considerable variation in the type of goals people might want to attain through the same activity. Whereas some individuals perceive an activity (e.g., exercising) as serving the attainment of an intrinsic goal (e.g., health), others might perceive the same activity as serving the attainment of an extrinsic goal (e.g., physical appearance). Therefore, we examined whether the effect of experimental intrinsic versus extrinsic goal framing would also hold for individuals who perceived the activity as serving an extrinsic goal. This is an important question, as it might be argued that most of the topics (e.g., healthy lifestyles, recycling) that participants learned about in previous studies on intrinsic versus extrinsic goal framing (e.g., Vansteenkiste, Simons, Lens, Sheldon, & Deci, 2004; Vansteenkiste, Simons, et al., 2005) had, on average, a rather intrinsic goal character. Hence, the observed positive effects of intrinsic goal framing when participants learned about these topics might have been due to the match between the induced intrinsic goal and the spontaneous perception of the activity as serving an intrinsic goal. The present study aimed to shed more light on this possible confound by directly examining whether the match between one's personal perception of the activity as serving an intrinsic or an extrinsic goal and the experimentally induced intrinsic versus extrinsic goal would promote learning, performance, and persistence.

### The Present Research

The present experimental field study aimed to further examine the conflicting hypotheses that can be derived from SDT and the match perspective. Participants in the current project were fifth- and sixth-grade children who learned more about a prosocial organization that is very well known in Flanders (Belgium)—that is, the Father Damian Foundation. This foundation financially supports individuals with tuberculosis. Children in the intrinsic goal condition were told that reading about and supporting the Father Damian Foundation could be important to help patients with tuberculosis (i.e., community contribution), whereas partici-

pants in the extrinsic goal condition were told that reading about and supporting the Father Damian Foundation could be important to make a good impression on their peers, parents, and teachers (i.e., social recognition).

One week prior to their participation in the field experiment, children provided answers to two sets of questions. First, they indicated to what extent they personally valued the intrinsic goal of community contribution and the extrinsic goal of social recognition. Second, they indicated whether, if they were asked to participate in activities that support the Father Damian Foundation, they would do so to make a good impression on others and to gain social recognition (extrinsic goal) or to help people with tuberculosis (intrinsic goal). Thus, prior to their participation in the actual experiment, children indicated whether they saw a potential engagement in the Father Damian Foundation, which was also the topic they would learn about during the experimental phase, as serving the attainment of an intrinsic or an extrinsic goal.

One week after we assessed participants' personal intrinsic versus extrinsic goal orientation and their intrinsic versus extrinsic goal perception of the activity, children were randomly assigned to an intrinsic or extrinsic goal-framing condition that either matched or mismatched their personal goal orientation and their personal task perception. Because the intrinsic and extrinsic goals prevailing in the environment were not subjectively assessed but instead were experimentally induced, the present research allowed for a direct examination of the effect of an objective person-environment match (Ostroff et al., 2005).

Dependent variables in the current project included the assessment of participants' autonomous motivation for engaging in the learning activity as well as their graded performance and free-choice persistence. *Autonomous motivation* refers to a willing or volitional engagement in learning, because the learning activity is perceived as interesting and enjoyable (intrinsic motivation) and as personally meaningful (identified motivation). Two types of performance were assessed—that is, conceptual learning, which requires deep and thoughtful processing of information and a more creative and integrative solution, and rote learning, which only requires a superficial engagement in the learning and has a more straightforward or rote path to the solution (literal memorization of factual information is sufficient; Entwistle & Entwistle, 1991; Grolnick & Ryan, 1987). Although past research has shown that extrinsic goal framing yields a debilitating effect on conceptual learning, such negative effects were not observed for rote learning (Vansteenkiste et al., 2005). Presumably, extrinsic goal framing results in an attentional shift away from the activity at hand, which precludes a thoughtful elaboration of the learning material, which is necessary for conceptual learning to take place.

The following three hypotheses were tested. First, on the basis of SDT, it was expected that intrinsic vs. extrinsic goal framing would predict more autonomous motivation, more conceptual (but not rote) learning, and more persistence for all learners—that is, for both intrinsic and extrinsic goal-oriented individuals. Second, on the basis of the match perspective, it was examined whether intrinsic goal-oriented individuals would benefit from intrinsic goal framing (i.e., match) and whether their learning would be undermined in the extrinsic goal condition (i.e., mismatch). Similarly, from this perspective, it can be anticipated that extrinsic goal-oriented individuals should benefit from extrinsic goal framing (i.e., match), whereas their learning should be forestalled in the



intrinsic goal condition (i.e., mismatch). Technically speaking, we examined whether the main effect of intrinsic versus extrinsic goal framing would be moderated by participants' personal intrinsic versus extrinsic goal orientation (i.e., interaction effect).

Third, on the basis of the match perspective, it was examined whether participants who saw the activity as serving an intrinsic goal would benefit from intrinsic goal framing (i.e., match), whereas their learning would be undermined in the extrinsic goal framing condition (i.e., mismatch). Similarly, from this perspective, it was expected that participants who saw the activity as serving an extrinsic goal would display more optimal learning when placed in the extrinsic goal framing condition (i.e., match), whereas their learning would be undermined in the intrinsic goal framing condition (i.e., mismatch). Accordingly, we examined whether the predicted main effect of contextual intrinsic versus extrinsic goal framing would be moderated by individuals' perception of the activity as serving an extrinsic or an intrinsic goal.

## Method

### *Participants and Procedure*

Seventy female and 68 male fifth- and sixth-grade children ( $N = 138$ ) participated in the study. In a first stage, which took place 1 week prior to the actual experiment, participants filled out questionnaires assessing their personal intrinsic versus extrinsic goal orientations and their perception of the intrinsic versus extrinsic goal character of the activity they would participate in about 1 week later (see below). The second stage took place during a religious meeting—preparing Roman Catholic children for their confirmation—in which participants were given a text about the Father Damian Foundation that they would have to read anyway.

The study took place in schools that are located in the town where Father Damian was born. Father Damian is a very well-known and popular public figure in this neighborhood, and several activities are organized yearly in this area to support the foundation and to remember Father Damian. For this reason, it is very normal that children at a relatively young age learn more about Father Damian and the Father Damian Foundation. The present experiment took place at the time the Damian campaign was organized.

Within each class, children were randomly assigned to an intrinsic or an extrinsic goal condition (cell sample sizes of 68 and 70, respectively) and read an instruction sheet (in Dutch) prior to studying the Father Damian text. We made sure that instruction sheets were of similar length so that participants would be unaware that they had received different instructions. The instruction sheets were turned in at the end of the session, after the children had written their names on them. The instructions included one of two different goal inductions for the forthcoming learning task. In the intrinsic going-framing condition, the children read that by learning more about the Father Damian Foundation, they could learn how to help individuals with tuberculosis (i.e., “Doing your best to read the text about the Father Damian Foundation might help you to know more about how you could help people with tuberculosis. By supporting the Father Damian Foundation you can help to save the lives of tuberculosis patients”), which constituted the intrinsic goal of community contribution (Kasser & Ryan, 1996). In the extrinsic goal-framing condition, they read that learning more

about and supporting the Father Damian Foundation would help them to attain the goal of being admired by others (i.e., “Doing your best to read the text about the Father Damian Foundation might help you to collect a lot of money and hence be admired by others. Thus, supporting the Father Damian Foundation is important to gain the social recognition of others”), which reflected the extrinsic goal of social recognition (Kasser & Ryan, 1996).

After studying the text, participants completed a questionnaire that assessed their degree of autonomous motivation for reading the text. Immediately following the completion of this questionnaire, participants were tested on their knowledge concerning the text (20 min). Subsequently, they were told that they could take home a comic book (i.e., drawing) about the Father Damian Foundation if they wanted to.

### *Measures*

*Personal intrinsic and extrinsic goal orientation.* Participants' intrinsic and extrinsic goal orientations are typically assessed with Kasser and Ryan's (1996) Aspiration Index. In previous research among Dutch-speaking university (Vansteenkiste, Duriez, et al., 2006) and high school students (Duriez, Vansteenkiste, Soenens, & De Witte, 2007), we have successfully developed a shortened Dutch version of this questionnaire. In the current study, participants filled out an adapted (i.e., simplified) child version of this shortened Dutch version of the Aspiration Index. They recorded how much they valued both types of goals by circling a number on a 5-point Likert scale ranging from 1 (*not important at all*) to 5 (*very important*). We only used the data for the intrinsic goal of community contribution (four items) and the extrinsic goal of social recognition (four items), as these represent the two goals that were manipulated in the present research.<sup>2</sup>

An exploratory factor analysis using promax rotation indicated that two factors needed to be retained (eigenvalues = 3.52 and 2.73), explaining 78% of the variance across the four community contribution items (e.g., “Helping other people”) and the four social recognition items (e.g., “I will be admired by other people”). Each of the items had a minimal loading of .34. We constructed social recognition and community contribution scores by averaging the social recognition and community contribution items, respectively. Internal consistencies of the community contribution and social recognition scores were .81 and .94, respectively. These analyses provide evidence for the structure of our goal measure and indicate that children's personal valuation of community contribution and social recognition can be reliably measured.

*Intrinsic and extrinsic task perception.* Intrinsic and extrinsic task perception were assessed with two 1-item measures. Partici-

<sup>2</sup> We believe that only using the goals of community contribution and social recognition, instead of the full Intrinsic and Extrinsic Goals scales, allows for a more accurate testing of the match hypothesis. This is because the match between the personally held goals (e.g., community contribution or social recognition) and the contextually promoted goal (e.g., community contribution or social recognition) would be strongest if one only retained the intrinsic goal of community contribution and the extrinsic goal of social recognition. The inclusion of the other intrinsic and extrinsic goals in the personal goal orientation measure would result in a poorer match between the personally held and contextually promoted goals and, as a result, would represent a methodologically weaker way of testing the match hypothesis.



pants were asked how important they found it to support the Father Damian Foundation to achieve the intrinsic goal of community contribution (i.e., “I prefer to support the Father Damian Project because it gives me the opportunity to help other people”) or the extrinsic goal of social recognition (i.e., “I prefer to support the Father Damian Project because of the good impression it makes on other people”). We focused on these two goals only because they represent the two goals that were manipulated during the actual experiment. On both items, participants indicated their agreement on a 10-point Likert scale that varied from 1 (*completely disagree*) to 10 (*completely agree*).

*Autonomous motivation.* Consistent with SDT and previous experimental studies among children (Vansteenkiste et al., 2005), autonomous motivation was measured by the extent to which participants found the reading material enjoyable and interesting (e.g., “I read the text because I found it very interesting”; intrinsic motivation; four items;  $\alpha = .91$ ) and personally relevant and meaningful (e.g., “I read this text because the content is personally meaningful to me”; identified regulation; four items;  $\alpha = .81$ ). Participants indicated their agreement with each of the items by circling a number between 1 (*completely disagree*) and 5 (*completely agree*). As in previous research (e.g., Vansteenkiste, Lens, Dewitte, De Witte, & Deci, 2004), we created a composite score of autonomous motivation by averaging the scores on these two subscales ( $\alpha = .82$ ).

*Test performance.* Two aspects of learning were assessed—that is, conceptual and rote learning. Five short questions tapped rote learning. To answer these questions, children needed to fill in one single word that was specifically mentioned in the text on Father Damian. As a result, scale scores for rote learning varied between 1 and 5. Concerning conceptual learning, the children were given three problems, which were scored on a scale varying from 1 (*very bad*) to 15 (*very good*). The first two conceptual problems were as follows: “Provide three examples that illustrate that Father Damian is a ‘stubborn stayer’” (4 points), and “Provide three reasons for the beatification of Father Damian” (3 points). The third problem was a crossword puzzle, which required children to fill in eight words (8 points). Each of the crossword questions tapped a deeper understanding of the learning material, as the answer to the question was not mentioned in the text but needed to be derived from it by meaningful linking of different pieces of information. These problems were created by the teacher and had been used in previous years to test children’s knowledge with respect to Father Damian. Therefore, the test was ecologically valid, and the children would have had to respond to the questions even if they had not been involved in the study (see Vansteenkiste et al., 2005, for a similar methodology). The regular teacher and one independent and trained rater, who were both blind to the nature of the study, determined whether the answers to these rote and conceptual learning questions were correct (1) or incorrect (0). There was perfect agreement between the raters scoring each question, as indexed by a perfect Pearson correlation between the two sets of rating scores.

*Persistence.* Participants were offered the opportunity to take home a comic book on Father Damian so that they could learn more about him and his foundation. The instructor registered who took home a comic book. In total, 75 participants (54%) did so. This constitutes a behavioral measure of free-choice persistence, as

it represents the tendency to display continued interest in the learning material once the learning task is terminated.

Results

Preliminary Analyses

One-sample *t* testing,  $t(137) = 7.90, p < .001$  indicated that, as a whole, the children valued the intrinsic goal of community contribution ( $M = 3.34, SD = 1.11$ ) more strongly than the extrinsic goal of social recognition ( $M = 2.43, SD = 0.88$ ), a finding that replicates previous work on intrinsic versus extrinsic goals among undergraduate students (e.g., Grouzet et al., 2006). To examine what percentage of children valued social recognition over community contribution and what percentage of children valued community contribution over social recognition, we subtracted participants’ social recognition score from their community contribution score. Then we dichotomized this difference measure by assigning all participants who had a negative difference score (indicating that they valued social recognition over community contribution) a score of  $-1$  and by assigning all participants who had a positive difference score (indicating that they valued community contribution over social recognition) a score of  $1$ . In total, 30% of the participants ( $n = 41$ ) valued social recognition over community contribution, and 70% of the participants ( $n = 97$ ) valued community contribution over social recognition. In the remainder of this article, the former group is referred to as the *extrinsic goal-oriented participants*, whereas the latter group is referred to as the *intrinsic goal-oriented participants*.

Next, we created four different groups by crossing contextual intrinsic versus extrinsic goal framing with the dichotomized personal intrinsic versus extrinsic goal orientation measure. Then we performed a chi-square test on the percentages of intrinsic and extrinsic goal-oriented individuals who were placed in the intrinsic and extrinsic goal-framing conditions. The purpose of this analysis was to examine whether intrinsic and extrinsic goal-oriented participants were equally distributed over the intrinsic and extrinsic goal-framing conditions. The percentages and exact cell sizes can be found in Table 1. The chi-square test turned out to be nonsignificant,  $\chi^2(1, N = 138) = 0.85$ , indicating that the intrinsic and extrinsic goal-framing conditions consisted of an equal percentage of intrinsic and extrinsic goal-oriented individuals.

Table 1  
Frequencies and Cell Sizes per Condition (Contextual Goal Framing × Personal Goal Orientation)

Contextual goal framing	Personal goal orientation		Row total
	Intrinsic	Extrinsic	
Intrinsic			
%	71	29	51
<i>n</i>	50	20	70
Extrinsic			
%	69	31	49
<i>n</i>	47	21	68
Column total			
%	70	30	100
<i>n</i>	97	41	138

A similar set of analyses was performed on participants' intrinsic versus extrinsic task perception. One-sample  $t$  testing,  $t(137) = 8.13$ ,  $p < .001$ , indicated that, on average, children were more likely to see their participation in the Father Damian Foundation as serving the intrinsic goal of community contribution ( $M = 6.54$ ,  $SD = 1.80$ ) than as serving the extrinsic goal of social recognition ( $M = 4.01$ ,  $SD = 2.23$ ). To examine what percentage of children perceived participation in the Father Damian Foundation as more strongly serving community contribution than social recognition and what percentage of children saw their participation as more strongly serving social recognition than community contribution, we subtracted the former scale from the latter. Then we dichotomized this continuous difference scale by assigning a score of  $-1$  if the participants had a negative difference score (indicating that they saw their participation as more strongly serving social recognition than community contribution) and assigning a score of  $1$  if participants had a positive difference score (indicating that they saw their participation as more strongly serving community contribution than social recognition). In total, 33% of the participants ( $n = 46$ ) saw their participation in the Father Damian Foundation as serving social recognition more strongly than community contribution, whereas 67% of the participants ( $n = 92$ ) saw their participation as more strongly serving community contribution.

Next, we created four groups by crossing intrinsic versus extrinsic goal framing with the constructed dichotomous intrinsic versus extrinsic task perception measure. Percentages and exact numbers of participants per group can be found in Table 2. Again, we performed a chi-square test with the purpose of examining whether individuals who perceived the task as intrinsically or extrinsically goal oriented were equally distributed across the two experimental conditions. The chi-square test turned out to be nonsignificant,  $\chi^2(1, N = 138) = 0.72$ , indicating that both groups of participants were equally distributed over the intrinsic and extrinsic goal-framing conditions.

In sum, these preliminary analyses indicated (a) that the sample of participants was quite heterogeneous in terms of both its intrinsic-extrinsic goal profile and its perception of the Father Damian Foundation as serving an intrinsic or extrinsic goal; (b) that the extrinsic and intrinsic goal-oriented participants, as well as the participants who saw the Father Damian Foundation as primarily serving an intrinsic goal or an extrinsic goal, were equally

distributed across the two experimental conditions; and (c) that each of the created groups, which represented a combination of contextual goal framing with either participants' personal goal orientation or their task perception, consisted of a sufficiently large number of participants (more than 20). Thus, these preliminary analyses indicated that we could proceed with the testing of our primary hypotheses.

Finally, to examine possible gender effects, we performed an independent-sample  $t$  test including autonomous motivation, performance, and persistence as dependent variables. Girls ( $M = 2.94$ ,  $SD = 0.48$ ) were found to be more autonomously motivated than boys ( $M = 2.75$ ,  $SD = 0.51$ ),  $t(136) = 2.28$ ,  $p < .05$ , but no gender differences were found with respect to conceptual learning, rote learning, and persistence. Hence, we controlled for gender effects in subsequent analyses when predicting autonomous motivation.

### Primary Analyses

The means and standard deviations for all variables across conditions, as well as the correlations between each pair of variables, are presented in Table 3. We examined our central hypotheses by performing two sets of regression analyses, one involving the difference score of intrinsic versus extrinsic goal orientations as a moderator, and one involving the difference score of intrinsic versus extrinsic activity perceptions as a moderator of experimental intrinsic versus extrinsic goal framing. We conducted multiple regression analyses to examine the effects on the continuous outcomes of autonomous motivation and both types of performance, but we made use of logistic regression analyses to examine the effects on the categorical outcome of persistence. In these primary analyses, we preferred to use the continuous (instead of the dichotomous) difference measures of intrinsic versus extrinsic goal orientations and intrinsic versus extrinsic goal task perception to increase statistical power.

In the first series of regressions, experimental intrinsic versus extrinsic goal framing, the continuous difference measure of intrinsic versus extrinsic goal orientation and the goal framing  $\times$  goal orientation interaction measure were entered in the regression equation. Experimental goal framing and goal orientation were centered, and the interaction term was calculated as the product of these centered scores. Multiple regression analyses indicated that experimental intrinsic, relative to extrinsic, goal framing positively predicted autonomous motivation ( $\beta = .43$ ,  $p < .001$ ) and conceptual learning ( $\beta = .31$ ,  $p < .001$ ) but was unrelated to rote learning ( $\beta = .13$ ,  $ns$ ). Intrinsic versus extrinsic goal orientation positively predicted autonomous motivation ( $\beta = .17$ ,  $p < .05$ ) but was unrelated to conceptual learning ( $\beta = .01$ ,  $ns$ ) and rote learning ( $\beta = -.05$ ,  $ns$ ).

On the basis of the match perspective, the main effect of experimental goal framing should be altered depending on individuals' personal goal orientations. However, the interactions between experimental goal framing and goal orientation in the prediction of autonomous motivation ( $\beta = -.08$ ,  $ns$ ), conceptual learning ( $\beta = -.01$ ,  $ns$ ), and rote learning ( $\beta = .06$ ,  $ns$ ) were not significant. Concerning persistence, a logistic regression analysis pointed out that the three predictors yielded an overall effect on persistence (dummy coded as lack of persistence = 0 and persistence = 1),  $\chi^2(3, N = 138) = 25.58$ ,  $p < .001$ . Persistence was

Table 2  
Frequencies and Cell Sizes per Condition (Contextual Goal Framing  $\times$  Goal Activity Perception)

Contextual goal framing	Goal activity perception		Row total
	Intrinsic	Extrinsic	
Intrinsic			
%	69	31	51
<i>n</i>	48	22	70
Extrinsic			
%	65	35	49
<i>n</i>	44	24	68
Columnn total			
%	67	33	100
<i>n</i>	92	46	138



Table 3  
Means, Standard Deviations, and Intercorrelations Between Independent and Dependent Variables

Variable	<i>M</i>	<i>SD</i>	Observed Range	1	2	3	4	5	6
1. Personal intrinsic vs. extrinsic goal	0.91	1.36	−3 to 4	—					
2. Intrinsic vs. extrinsic task perception	2.53	3.66	−6 to 8	.60**	—				
3. Autonomous motivation	2.92	0.50	2 to 4	.16*	.15	—			
4. Conceptual learning	9.41	2.39	3 to 15	.02	.11	.22*	—		
5. Rote learning	3.78	0.81	2 to 5	−.06	.14	.08	.45**	—	
6. Persistence	0.54	0.50	0 to 1	.10	.22*	.22*	.47**	.32**	—

\*  $p < .05$ . \*\*  $p < .01$ .

significantly predicted by placement in the experimental intrinsic goal-framing condition (odds ratio [OR] = 0.17,  $p < .001$ ) but not by the personal goal orientation measure (OR = 1.22, *ns*) nor by the interaction between experimental goal framing and the personal goal orientation measure (OR = 0.94, *ns*). Together, the predictors explained between 2% and 26% of the variance in the outcomes.

In a second set of regressions, the three outcomes were regressed on experimental intrinsic versus extrinsic goal framing, the continuous difference measure of intrinsic versus extrinsic activity perception, and the interaction between experimental goal framing and activity perception. Concerning autonomous motivation and performance outcomes, multiple regression analyses indicated that experimental intrinsic versus extrinsic goal framing positively predicted autonomous motivation ( $\beta = .42$ ,  $p < .001$ ) and conceptual learning ( $\beta = .34$ ,  $p < .001$ ) but was unrelated to rote learning ( $\beta = .09$ ). Intrinsic versus extrinsic activity perception was unrelated to autonomous motivation ( $\beta = .11$ , *ns*), conceptual learning ( $\beta = .10$ , *ns*), and rote learning ( $\beta = .13$ , *ns*). Most important, the interaction between experimental goal framing and activity perception did not reach significance (autonomous motivation,  $\beta = .14$ , *ns*; conceptual learning,  $\beta = -.05$ , *ns*; and rote learning,  $\beta = .05$ , *ns*). Concerning persistence, a logistic regression analysis indicated that the three predictors yielded an overall effect,  $\chi^2(3, N = 138) = 30.69$ ,  $p < .001$ . Persistence was significantly predicted by placement in the experimental intrinsic goal framing condition (OR = 0.14,  $p < .001$ ) and by the presence of an intrinsic activity perception (OR = 1.15,  $p < .01$ ) but not by the interaction between experimental goal framing and the activity perception measure (OR = 0.93, *ns*). Together, the predictors explained between 4% and 27% of the variance in the outcomes.<sup>3</sup>

## Discussion

The present research examined conflicting hypotheses derived from SDT and the match perspective regarding the impact of intrinsic versus extrinsic goal framing on learning and performance. This research extends previous work in this area by using an experimental instead of a correlational design, by focusing on learning and performance instead of well-being as outcomes, and by examining the possible moderating role of two types of individual-differences variables (i.e., intrinsic vs. extrinsic goal orientation and intrinsic vs. extrinsic activity perception) in the relation between intrinsic versus extrinsic goal framing and learning and performance. Several important findings emerged.

Several previous studies among fifth- to sixth-grade children, high school students, and college students have shown that the experimental induction of intrinsic goals promotes better conceptual learning and persistence compared to the induction of extrinsic goals (see Vansteenkiste, Lens, & Deci, 2006). The present research fully replicated previous studies by showing that extrinsic, relative to intrinsic, goal framing undermined participants' self-reported task enjoyment and personal valuation of the activity (i.e., autonomous motivation) and forestalled their conceptual performance and persistence. The reason for these effects is probably that inducing an extrinsic goal creates an outward orientation by drawing individuals' attention toward the external signs of worth. This attentional shift precludes the possibility of task-absorbed engagement in the activity, so that participants fail to enjoy the learning activity and perform less well (Vansteenkiste, Lens, & Deci, 2006). As in previous experimental work (Vansteenkiste et al., 2005), the negative effect of extrinsic goal framing did not emerge for rote learning. Apparently, extrinsic goal framing prompts some motivated learning; however, the learning activity is approached in a more rigid and narrow-minded fashion, because the learning is primarily oriented toward attaining the extrinsic goals. Such a rigid and restricted approach is sufficient to promote memorization of learning material but interferes with an in-depth processing of the contents.

Beyond replicating past results, the main contribution of the present work was to examine whether the effects of contextual goal framing depended on participants' own goal preferences. Vansteenkiste, Simons, Lens, Sheldon, and Deci (2004) found that extrinsic (vs. intrinsic) goal framing undermined optimal learning among a group of business students, but they did not include assessments of individuals' own goal orientations prior to participation in the experiment. As such, the authors had no opportunity to directly examine whether goal framing interacted with personal goal orientations. In particular, research to date has not directly tested the possibility that the observed negative effects of extrinsic goal framing would be reversed for extrinsic goal-oriented indi-

<sup>3</sup> As recommended by one of the anonymous reviewers, we conducted a power analysis. A post hoc power analysis indicated that with the current sample size of 138,  $\alpha = .05$ , three used predictors, and  $R^2 = .20$ , the current set of regression analyses yielded sufficient power (i.e., 1.00) to detect significant effects. An a priori power analysis indicated that with a desired power of .80, an alpha level of .05, an estimated effect size of .25, and a total of three predictors, we would need to sample at least 48 participants. The current sample size meets this requirement amply.

viduals, as would be predicted by the match perspective. The present research addressed exactly this question.

On average, participants in the present sample were found to more strongly value the intrinsic goal of community contribution than the extrinsic goal of social recognition, a finding that is in line with various previous studies (e.g., Grouzet et al., 2006). In spite of this, 30% of the present sample attached higher importance to social recognition than to community contribution, which suggests that the current sample was relatively heterogeneous in terms of its intrinsic–extrinsic goal profile. Furthermore, the extrinsic goal-oriented individuals appeared to be equally distributed across the intrinsic and extrinsic goal-framing conditions, and, thus, participants appeared to be randomly dispersed across the experimental conditions. Interaction analyses between contextual goal framing and personal goal orientation indicated that the effect of contextual intrinsic versus extrinsic goal framing did not depend on participants' personal intrinsic versus extrinsic goal orientation. Thus, the match perspective could not be confirmed, as both intrinsic and extrinsic goal-oriented individuals benefited from being placed in an intrinsic goal condition. Presumably, the shift in attention that we caused by inducing an extrinsic goal negatively affected both intrinsic and extrinsic goal-oriented individuals. Referring to an extrinsic goal elicits self-worth concerns and distracts all participants' attention from the activity at hand, which undermines optimal learning (Vansteenkiste, Lens, & Deci, 2006). In contrast, learning in the service of intrinsic goals is more consistent with people's basic psychological needs, which should promote more optimal learning for all individuals (Deci & Ryan, 2000; Kasser, 2002).

The match perspective was also tested in a second way. We suggest that there exists considerable variability in the type of goals the same activity might serve. Whereas some individuals might perceive an activity as serving the attainment of intrinsic goals, other individuals would see the same activity as a means toward extrinsic goal attainment. A possible confound in previous work on intrinsic versus extrinsic goal framing is that the selected activities (e.g., reading a text on recycling) might, on average, be perceived as serving an intrinsic goal. When individuals are subsequently instructed that the activity serves an intrinsic goal, they find themselves in a match situation, whereas individuals who are told that the activity serves an extrinsic goal find themselves in a mismatch situation. The induced extrinsic goal might even be seen as meaningless and the instructions as unbelievable, because they conflict with participants' spontaneous goal perception of the activity.

To eliminate this possible confound, we obtained direct assessments of the type of goals individuals pursued through supporting the Father Damian Foundation, the topic they would read about during the experimental phase. Although children saw the activity, on average, as standing more strongly in the service of intrinsic than extrinsic goal attainment, there was substantial variation in their perceptions. That is, one third of the children saw the activity as instrumental for extrinsic goal attainment (i.e., social recognition), whereas the other two thirds saw the activity as instrumental for intrinsic goal attainment (i.e., community contribution). We found that the effect of intrinsic versus extrinsic goal framing was not altered for individuals who spontaneously saw the activity as serving an extrinsic goal. All individuals, regardless of their activity perception, displayed more optimal learning in the intrinsic, compared to the extrinsic, goal-framing condition.

Another set of findings that deserves some discussion is the observation that individuals' intrinsic versus extrinsic goal profiles and intrinsic versus extrinsic activity perception had smaller effects on their motivation, learning, and performance compared to the effects of contextual goal framing. Although these results need to be replicated in future research, we provide two explanations. First, the individual goal measures were assessed 1 week prior to children's participation in the experiment, whereas children read the instructions on intrinsic versus extrinsic goal framing a few minutes before reading the text. Because of this different time lag, it seems logical that contextual goal framing yielded stronger effects compared to the two individual-differences measures.

A second interpretation deals with the level of generality at which the independent variables can be situated. That is, within the hierarchical model of motivation (Vallerand, 1997), a global (i.e., personality), a domain-bounded (i.e., in a particular life domain), and a situational (i.e., for a specific task) level of motivation are distinguished. Clearly, the present study examined participants' learning and performance on a specific task—that is, at the situational level of motivation. Whereas participants' intrinsic and extrinsic goal profiles reflect their global endorsement of goals, contextual goal framing represents a manipulation at the situational level, because the presented goals are directly linked to the learning activity itself. Hence, it seems logical that contextual goal framing is a better predictor of situation-bounded outcomes than participants' goal profiles, because the former are situated at a similar level of motivation, whereas the latter are not (see also Roberts & Pomerantz, 2004, for a similar reasoning).

Does the current set of results imply that all types of goal-match situations will invariantly fail to predict optimal performance? We do not think so. Researchers have focused on very different types of goals and psychological constructs to examine the match hypothesis (see Amiot et al., 2006; Bianco et al., 2003; Harackiewicz & Elliot, 1998). The present research differs from that work because it focuses on the concepts of intrinsic versus extrinsic goals. To address the question of whether particular goals will promote optimal functioning, SDT suggests that researchers evaluate the functional significance of these goals by asking how they are dynamically related to basic psychological need satisfaction. If the pursuit or induction of particular goals fails to promote basic need satisfaction, a match situation is unlikely to override the negative effects associated with thwarted need satisfaction. Because extrinsic goal induction is, on average, unrelated or even negatively related to basic need satisfaction, these goals should result in diminished learning, even when individuals strongly value these goals, as convincingly shown in the present research. SDT's position regarding the motivational impact of goals also implies that if particular goals cannot be meaningfully tied to basic need satisfaction, placing the individual in a match situation might predict more optimal functioning. Overall, we believe that more research is needed to carefully examine whether the match hypothesis holds for particular goals, whereas it does not for other types of goals.

### *Practical Implications*

Previous research (e.g., Kasser & Ahuvia, 2002; Sagiv & Schwartz, 2000) investigated whether extrinsically oriented individuals benefit psychologically if their personal goals match with the goals emphasized in their environment. Accordingly, these studies focused on the question of whether a person's goals should



match with the environmentally promoted goals to predict psychological well-being (i.e., person–environment match). The present study asked a different question: Do socializing agents need to adjust their goal framing in accordance with the goal profiles of the individuals who need to be motivated? Put differently, does the environment need to be accommodated to the persons who are entering it (environment–person match)? If this were true, it would imply that socializing agents first need to gain insight into individuals' personal goal orientations and, subsequently, need to use individualized goal-framing instructions that match the person's goal preference to motivate learning. This seems an extremely difficult task. Luckily, the present experiment indicates that socializing agents' motivational task is much easier: It suffices to induce intrinsic goals when one is trying to promote optimal learning, even if individuals ascribe high importance to extrinsic goals themselves.

Of course, there might be a few boundary conditions to this goal framing effect that could be more thoroughly examined in future research. First, consistent with previous goal research (Locke & Latham, 1990, 2002), we suggest that a specific, instead of a vague, intrinsic goal is more likely to promote optimal learning (see also Vansteenkiste, Simons, Soenens, & Lens, 2004). Second, the induced intrinsic goal needs to be realistically and meaningfully connected to the learning activity, so that learners accept the provided intrinsic goal. Past research has shown that providing a goal that is not accepted by participants does not yield the same performance benefits (Locke & Latham, 1990). Third, although extrinsic goal framing undermines deep-level learning, the current study shows that extrinsic goal framing predicts a similar extent of superficial and shallow learning as intrinsic goal framing (Vansteenkiste et al., 2005).

### Limitations and Future Research

The current project contained a number of shortcomings that might be overcome in future research. First, only one single intrinsic goal and one single extrinsic goal were manipulated. In future studies, researchers might want to examine whether manipulation of other extrinsic goals (e.g., physical appearance, financial success) promotes or undermines optimal learning among individuals who attach importance to these extrinsic goals. Second, it would be interesting to develop a measure of experienced goal conflict versus goal congruence as a check on whether individuals placed in a mismatch condition effectively experience goal conflict, while those placed in a match condition experience goal congruence. Third, future research might sample extrinsically oriented individuals who are older (e.g., business college students) than the ones who participated in the current project to examine whether the match hypothesis could be confirmed (or not) among these older age groups. Finally, the current results might also be examined in different domains from learning and education, including sports and exercising, work, and cross-cultural research. For instance, it would be interesting to examine whether the pursuit of extrinsic goals yields different effects depending on whether one perceives one's coach, one's manager, the organization, or even the culture at large as promoting and reinforcing such goals.

### Conclusion

The results of the present study indicate that framing a learning activity in terms of the attainment of an extrinsic rather than an intrinsic goal undermined task enjoyment, personal valuation of a learning activity, conceptual learning, and persistence, irrespective of individuals' own intrinsic and extrinsic goals or their intrinsic and extrinsic task perception. Hence, these results suggest that instructors, teachers, and other socializing agents might do well to promote intrinsic goals and might downplay the promotion of extrinsic goals to motivate conceptual learning and performance, even for individuals who ascribe high importance to such extrinsic goals.

### References

- Amiot, C. E., Vallerand, R. J., & Blanchard, C. M. (2006). Passion and psychological adjustment: A test of the person-environment fit hypothesis. *Personality and Social Psychology Bulletin*, 32, 220–229.
- Bianco, A. M., Higgins, E. T., & Klem, A. (2003). How “fun/importance” fit affects performance: Relating implicit theories to instructions. *Personality and Social Psychology Bulletin*, 29, 1091–1103.
- Cohen, P., & Cohen, J. (1996). *Life values and adolescent mental health*. Mahwah, NJ: Erlbaum.
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum Press.
- Deci, E. L., & Ryan, R. M. (2000). The “what” and the “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11, 227–268.
- De Witte, H. (2004). Ideological orientations and values. In C. Spielberger (Ed.), *Encyclopedia of applied psychology* (Vol. 2, pp. 1–10). Amsterdam: Elsevier.
- Duriez, B., Vansteenkiste, M., Soenens, B., & De Witte, H. (2007). The social costs of extrinsic relative to intrinsic goal pursuits: Their relation with social dominance and racial and ethnic prejudice. *Journal of Personality*, 75, 757–782.
- Eccles, J. S., Lord, S., & Midgley, C. (1991). What are we doing to early adolescents? The impact of educational contexts on early adolescents. *American Journal of Education*, 99, 521–541.
- Edwards, J. R. (1991). Person-job fit: A conceptual integration, literature review, and methodological critique. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology* (pp. 283–357). Oxford, England: Wiley.
- Entwistle, N., & Entwistle, A. (1991). Contrasting forms of understanding for degree examinations: The student perspective and its implications. *Higher Education*, 22, 205–277.
- Fromm, E. (1976). *To have or to be?* New York: Continuum.
- Grolnick, W. S., & Ryan, R. M. (1987). Autonomy in children's learning: An experimental and individual difference investigation. *Journal of Personality and Social Psychology*, 52, 890–898.
- Grouzet, F. M. E., Kasser, T., Ahuvia, A., Dols, J. M. F., Kim, Y., Lau, S., et al. (2006). The structure of goal contents across 15 cultures. *Journal of Personality and Social Psychology*, 89, 800–816.
- Harackiewicz, J. M., & Elliot, A. J. (1998). The joint effects of target and purpose goals on intrinsic motivation: A mediational analysis. *Personality and Social Psychology Bulletin*, 24, 675–689.
- Holland, J. L. (1985). *Making vocational choices: A theory of careers*. Englewood Cliffs, NJ: Prentice-Hall.
- Kasser, T. (2002). *The high price of materialism*. Cambridge, MA: MIT Press.
- Kasser, T. (2005). Frugality, generosity, and materialism for use in children and adolescents. In K. A. Moore & L. H. Lippman (Eds.), *What do children need to flourish? Conceptualizing and measuring indicators of positive development* (pp. 357–374). New York: Springer.

- Kasser, T., & Ahuvia, A. (2002). Materialistic values and well-being in business students. *European Journal of Social Psychology*, 32, 137–146.
- Kasser, T., & Ryan, R. M. (1993). A dark side of the American dream: Correlates of financial success as a central life aspiration. *Journal of Personality and Social Psychology*, 65, 410–422.
- Kasser, T., & Ryan, R. M. (1996). Further examining the American dream: Differential correlates of intrinsic and extrinsic goals. *Personality and Social Psychology Bulletin*, 22, 280–287.
- Kasser, T., Ryan, R. M., Couchman, C. E., & Sheldon, K. (2004). Materialistic values: Their causes and consequences. In T. Kasser & A. D. Kanner (Eds.), *Psychology and consumer culture: The struggle for a good life in a materialistic world* (pp. 11–28). Washington, DC: American Psychological Association.
- Lepper, M. R., & Greene, D. (1978). *The hidden costs of reward*. Hillsdale, NJ: Erlbaum.
- Locke, E. A., & Latham, G. P. (1990). Work motivation and satisfaction: Light at the end of the tunnel. *Psychological Science*, 1, 240–245.
- Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation. *American Psychologist*, 57, 705–717.
- Maslow, A. H. (1954). *Motivation and personality*. New York: Harper.
- Meglino, B. H., Ravlin, E. C., & Adkins, C. L. (1989). A work value approach to corporate culture: A field test of the value congruence process and its relationship to individual outcomes. *Journal of Applied Psychology*, 74, 424–432.
- Nuttin, J. R., & Lens, W. (1985). *Future time perspective and motivation: Theory and research method*. Leuven, Belgium/Hillsdale, NJ: Leuven University Press/Erlbaum.
- Ostroff, C. (1993). The effects of climate and personal influences on individual behavior and attitudes in organizations. *Organizational Behavior and Human Decision Processes*, 56, 56–90.
- Ostroff, C., Shin, Y., & Kinicki, A. J. (2005). Multiple perspectives of congruence: Relationships between value congruence and employee attitudes. *Journal of Organizational Behavior*, 26, 591–623.
- Pervin, L. A. (1968). Performance and satisfaction as a function of individual–environment fit. *Psychological Bulletin*, 69, 56–68.
- Roberts, B. W., & Pomerantz, E. M. (2004). On traits, situations, and their integration: A developmental perspective. *Personality and Social Psychology Review*, 8, 402–416.
- Rohan, M. J. (2000). A rose by a name? The values construct. *Personality and Social Psychology Review*, 4, 255–277.
- Rokeach, M. (1973). *The nature of human values*. New York: Free Press.
- Ryan, R. M., & Deci, E. L. (2000a). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25, 54–67.
- Ryan, R. M., & Deci, E. L. (2000b). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55, 68–78.
- Ryan, R. M., Mims, V., & Koestner, R. (1983). Relation of reward contingency and interpersonal context to intrinsic motivation: A review and test using cognitive evaluation theory. *Journal of Personality and Social Psychology*, 45, 736–750.
- Ryan, R. M., Sheldon, K. M., Kasser, T., & Deci, E. L. (1996). All goals were not created equal: An organismic perspective on the nature of goals and their regulation. In P. M. Gollwitzer & J. A. Bargh (Eds.), *The psychology of action: Linking motivation and cognition to behavior* (pp. 7–26). New York: Guilford Press.
- Sagiv, L., & Schwartz, S. H. (2000). Value priorities and subjective well-being: Direct relations and congruity effects. *European Journal of Social Psychology*, 30, 177–198.
- Schneider, B. (1987).  $E = f(P, B)$ : The road to a radical approach to person–environment fit. *Journal of Vocational Behavior*, 31, 353–361.
- Schwartz, S. H. (1992). Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 25, pp. 1–65). New York: Academic Press.
- Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values? *Journal of Social Issues*, 50, 19–45.
- Schwartz, S. H. (1999). A theory of cultural values and some implications for work. *Applied Psychology: An International Review*, 48, 23–47.
- Srivastava, A., Locke, E. A., & Bartol, K. M. (2001). Money and subjective well-being: It's not the money, it's the motives. *Journal of Personality and Social Psychology*, 80, 959–971.
- Vallerand, R. (1997). Toward a hierarchical model of intrinsic and extrinsic motivation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 29, pp. 271–360). San Diego, CA: Academic Press.
- Vansteenkiste, M., Duriez, B., Simons, J., & Soenens, B. (2006). Materialistic values and well-being among business students: Further evidence for their detrimental effect. *Journal of Applied Social Psychology*, 36, 2892–2908.
- Vansteenkiste, M., Lens, W., & Deci, E. L. (2006). Intrinsic versus extrinsic goal contents in self-determination theory: Another look at the quality of academic motivation. *Educational Psychologist*, 41, 19–31.
- Vansteenkiste, M., Lens, W., Dewitte, S., De Witte, H., & Deci, E. L. (2004). The “why” and “why not” of job search behavior: Their relation to searching, unemployment experience, and well-being. *European Journal of Social Psychology*, 34, 345–363.
- Vansteenkiste, M., Neyrinck, B., Niemiec, C. P., Soenens, B., De Witte, H., & Van den Broeck, A. (2007). On the relations among work value orientations, psychological need satisfaction, and job outcomes: A self-determination theory approach. *Journal of Occupational and Organizational Psychology*, 80, 251–277.
- Vansteenkiste, M., Ryan, R. M., & Deci, E. L. (in press). Self-determination theory and the explanatory role of psychological needs in human well-being. In L. Bruni, F. Comim, & M. Pugno (Eds.), *Capabilities and happiness*. Oxford, England: Oxford University Press.
- Vansteenkiste, M., Simons, J., Lens, W., Sheldon, K. M., & Deci, E. L. (2004). Motivation learning, performance and persistence: The synergistic effects of intrinsic goal contents and autonomy-supportive contexts. *Journal of Personality and Social Psychology*, 87, 246–260.
- Vansteenkiste, M., Simons, J., Lens, W., Soenens, B., & Matos, L. (2005). Examining the impact of extrinsic versus intrinsic goal framing and internally controlling versus autonomy-supportive communication style on early adolescents' academic achievement. *Child Development*, 76, 483–501.
- Vansteenkiste, M., Simons, J., Lens, W., Soenens, B., Matos, L., & Lacante, M. (2004). Less is sometimes more: Goal content matters. *Journal of Educational Psychology*, 96, 755–764.
- Vansteenkiste, M., Simons, J., Soenens, B., & Lens, W. (2004). How to become a persevering exerciser? Providing a clear, future intrinsic goal in an autonomy supportive way. *Journal of Sport & Exercise Psychology*, 26, 232–249.
- Vansteenkiste, M., Soenens, B., & Lens, W. (2007). Intrinsic versus extrinsic goal promotion in exercise and sport: Understanding the differential impacts on performance and persistence. In M. S. Hagger & N. Chatzisarantis (Eds.), *Intrinsic motivation and self-determination in exercise and sport* (pp. 167–180). Champaign, IL: Human Kinetics.
- Walsh, W., & Holland, J. L. (1992). A theory of personality types and work environments. In W. B. Walsh, K. H. Cracick, & R. H. Price (Eds.), *Person–environment psychology: Models and perspectives* (pp. 35–78). Hillsdale, NJ: Erlbaum.
- Williams, G. C., Cox, E. M., Hedberg, V. A., & Deci, E. L. (2000). Extrinsic life goals and health-risk behaviors in adolescents. *Journal of Applied Social Psychology*, 30, 1756–1771.

Received May 25, 2006

Revision received March 15, 2007

Accepted March 18, 2007 ■



# Task Values, Achievement Goals, and Interest: An Integrative Analysis

Chris S. Hulleman  
University of Wisconsin—Madison

Amanda M. Durik  
Northern Illinois University

Shaun A. Schweigert and Judith M. Harackiewicz  
University of Wisconsin—Madison

The research presented in this article integrates 3 theoretical perspectives in the field of motivation: expectancy-value, achievement goals, and interest. The authors examined the antecedents (initial interest, achievement goals) and consequences (interest, performance) of task value judgments in 2 learning contexts: a college classroom and a high school sports camp. The pattern of findings was consistent across both learning contexts. Initial interest and mastery goals predicted subsequent interest, and task values mediated these relationships. Performance-approach goals and utility value predicted actual performance as indexed by final course grade (classroom) and coach ratings of performance (sports camp). Implications for theories of motivation are discussed.

*Keywords:* utility value, achievement goals, interest, expectancy value, performance

The definition of optimal motivation . . . is motivation that produces optimum intellectual development. Though optimum intellectual development is the ultimate goal, progress toward it is assessed in terms of motivation. (Nicholls, 1979, p. 1072)

As Nicholls (1979) asserted, motivation and achievement are inherently connected. An understanding of the motivational dynamics at work in achievement settings will therefore allow researchers and educators to better understand how to promote learning. Motivation is defined here as a motive (e.g., wish, intention, drive) to engage in a specific activity (Austin & Vancouver, 1996; Schiefele, 1999; Weiner, 1985), and can be conceptualized as a behavioral antecedent, a process experienced during task engagement, and as an outcome. Utilizing this multifaceted conceptualization of motivation enables an integration of three distinct, yet overlapping, theoretical perspectives: expectancy-value (Eccles, 2005), achievement goal (Dweck, 1986; Nicholls, 1984), and interest theory (Hidi & Renninger, 2006). These three per-

spectives allow researchers to understand motivational dynamics in a way that any one single perspective may not completely capture.

Interest in activities has been considered to be one of the central components of motivation and motivated behavior (Deci & Ryan, 1985; Dewey, 1913; Schiefele, 1991). One way to develop interest in activities is to find meaning and value in those activities (Hidi & Renninger, 2006; Renninger & Hidi, 2002). Achievement goals and initial interest can predispose individuals to find value in educational activities (Hidi & Harackiewicz, 2000; Pintrich, 2003; Wigfield & Eccles, 2002). Thus, we propose a model of motivation wherein achievement goals and initial interest lead to the perception of task value, which then promotes the development of subsequent interest and learning. In the present article, we examine the antecedents and consequences of task value judgments in two learning contexts: a college classroom and a high school sports camp.

## Task Values

Research from the expectancy-value perspective has examined the values that individuals perceive when engaging in tasks, and how these task values are related to subsequent achievement choices. Eccles and her colleagues (Eccles et al., 1983; see Eccles, 2005, for a review) have identified several types of task value that are important in predicting motivation and achievement, in addition to the well-documented effects of success expectancies (Bandura, 1997; Pintrich & Schunk, 2002). Two of these task values are utility and intrinsic value. Tasks with utility value are important because they are useful and relevant beyond the immediate situation, for other tasks or aspects of a person's life. Tasks with intrinsic value are important to the individual because they are enjoyable and fun. Individuals can discover and appreciate the value of activities through interaction and experience. Perceiving utility and/or intrinsic value in tasks has been associated with motivation and interest in activities. For example, both intrinsic

---

Chris S. Hulleman, Shaun A. Schweigert, and Judith M. Harackiewicz, Department of Psychology, University of Wisconsin—Madison; Amanda M. Durik, Department of Psychology, Northern Illinois University.

Chris S. Hulleman is now at the Department of Psychology and Human Development, Peabody College, Vanderbilt University.

A previous version of this article was presented at the annual meeting of the American Educational Research Association, Montreal, Ontario, Canada, April 2005. This research was supported in part by grants from the Hilldale Undergraduate/Faculty Research Fellowship, the Department of Psychology Research Award, and the Interdisciplinary Training Program in Education Sciences Fellowship, all from the University of Wisconsin—Madison. We thank Jonathan Trinastic and Rich Kacmarynski for their assistance with data collection.

Correspondence concerning this article should be addressed to Chris S. Hulleman, who is now at 230 Appleton Place, Peabody #512, Vanderbilt University, Nashville, TN 37203-5721. E-mail: chris.hulleman@vanderbilt.edu

and utility value have been found to predict motivational outcomes such as course enrollment decisions (Harackiewicz, Durik, Barron, Linnenbrink-Garcia, & Tauer, 2008; Meece, Wigfield, & Eccles, 1990; Updegraff, Eccles, Barber, & O'Brien, 1996; Wigfield, 1994), self-reported effort in science classes (Cole, Bergin, & Whittaker, 2006; Mac Iver, Stipek, & Daniels, 1991), intentions to continue a school-based running program (Xiang, Chen, & Bruene, 2005), amount of free time spent on sports (Eccles & Harold, 1991), and classroom interest (Durik, 2003).

Whereas both intrinsic and utility task values have been linked to motivation, utility value may be uniquely related to achievement. For example, Simons, Dewitte, and Lens (2003) found that highlighting the usefulness of an activity—by telling participants how it could help them achieve their future goals—increased persistence and performance in a physical education class. Bong (2001b) found that the perceived usefulness of a course predicted self-efficacy in the course, which in turn predicted exam performance. Malka and Covington (2005) found that the relevance of school work to students' future goals (i.e., perceived instrumentality) predicted classroom performance. These studies indicate that there is a relationship between perceiving utility in a task and subsequent performance.

Given the role of intrinsic and utility value in predicting interest and performance, it is essential to understand how individuals come to perceive tasks as intrinsically valuable and useful. The expectancy-value model proposes that goals lead to task values, which are considered to be the more proximal predictors of achievement choices (Eccles et al., 1983; Wigfield & Eccles, 1992, 2002). A large body of research suggests that the achievement goals that students and athletes pursue serve to orient them to tasks (Duda, 1995), influence their cognitive processes and effort during task engagement (Linnenbrink & Pintrich, 2000), and help them find more or less value in what they do (Pintrich, 2003; Wigfield & Eccles, 1992). For example, a student whose goal is to learn and understand course material may be more likely to experience the intrinsic value of the material and see how the course is relevant to his or her life. This enjoyment and personal connection with the material may facilitate attention, cognitive processing, effort, and subsequent interest (Hidi, 1990; Hidi & Harackiewicz, 2000).

### Achievement Goals

According to the achievement goal perspective (Ames, 1984; Dweck, 1986; Nicholls, 1984), goals are mental representations of individuals' competence strivings during achievement activities (Maehr, 1989; Shah & Kruglanski, 2000). Although different labels have been used by researchers over the years, achievement goals have been divided into two general classes: mastery and performance (Ames & Archer, 1988), while also being subdivided into approach and avoidance components (Elliot, 1999; Elliot & McGregor, 2001; Pintrich, 2000a). In this article, we will focus on approach goals because we believe that approach motivation is the most relevant for the discovery of value and development of interest. Mastery-approach goals focus on developing knowledge and learning new skills, whereas performance-approach goals focus on doing well compared with other people. Research has demonstrated that mastery-approach goals positively predict a variety of motivational variables including classroom interest (Ames & Archer, 1988; Harackiewicz, Barron, Pintrich, Elliot, &

Thrash, 2002; Lee, Sheldon, & Turban, 2003), as well as deep processing, effort, and persistence (Elliot, McGregor, & Gable, 1999; Harackiewicz, Barron, Tauer, Carter, & Elliot, 2000; Midgley, Kaplan, & Middleton, 2001; Wolters, Yu, & Pintrich, 1996). On the other hand, performance-approach goals are positively linked to academic performance (Harackiewicz, Barron, Pintrich, et al., 2002; but see Grant & Dweck, 2003).

When individuals pursue performance goals, they focus on the outcome of task engagement (i.e., success/failure) and may not become deeply engaged in the activity. In contrast, individuals who pursue mastery goals focus on learning and improvement, and are able to experience and explore the activity because they are focused on the process rather than on the product of task engagement (Dewey, 1913; Flum & Kaplan, 2005; Renninger & Hidi, 2002). This task focus may allow them to discover or experience both intrinsic and utility value, which may then motivate effort, persistence, and subsequent interest. From this perspective, achievement goals operate as a framework for the perception of task value, and perceived task value becomes the more proximal predictor of interest and performance. Experimental studies have documented the effects of mastery goals on task involvement and task interest (Senko & Harackiewicz, 2005; Zimmerman & Kitsantas, 1997; for a review, see Rawsthorne & Elliot, 1999). Other research has demonstrated that mastery-approach goals are correlated with both utility and intrinsic value (Bong, 2001a; DeBacker & Nelson, 1999; Linnenbrink, 2005; Mac Iver et al., 1991; Shim & Ryan, 2005; Simons, Dewitte, & Lens, 2004; Wigfield & Eccles, 2002; Xiang, McBride, & Bruene, 2004).

In summary, our hypothesized model suggests that mastery goals create a framework within which individuals focus on the task and perceive value in it. Perceiving both intrinsic and utility value can lead to the development of subsequent interest. Although mastery goals have not been consistently associated with achievement (Harackiewicz, Barron, Pintrich, et al., 2002), mastery goal adoption may have indirect effects on performance through perceptions of utility value. Perceived utility value may emerge from deep engagement in an activity (facilitated by a mastery goal), and may motivate attention, effort, and persistence that result in higher levels of performance.

### The Development of Interest

How does the perception of value influence the development of interest? Interest can be defined as a psychological state, for example, "being engaged, engrossed, or entirely taken up with some activity" (Dewey, 1913, p. 17), and as a process that emerges over time (Hidi & Renninger, 2006). The psychological state of interest that arises through interactions with task features, such as stimulating pictures and humor, is known as *situational interest* (Hidi, 1990). Interest that is more enduring and traitlike, and that develops and deepens over time, is known as *individual interest* (Renninger, 2000). Perceiving value in activities has been hypothesized to be a key contributor in the progression from situational to individual interest, and to the deepening of existing individual interest (Hidi & Renninger, 2006). As momentary interest in the situation is activated, an individual may come to perceive value in the activity and desire to continue pursuing the activity over time, resulting in maintained situational interest. If situational interest is maintained and an individual continues to engage in the activity



and perceive value, then individual interest may develop (emerging individual interest). Thus, perceived value may play a critical role in the beginning stages of interest development as well as in the deepening of individual interest over time.

### Integration of Motivational Perspectives

Although we have conceptualized value and interest as distinct constructs, many theorists consider value to be an integral component of interest (Dewey, 1913; Hidi & Harackiewicz, 2000; Hidi & Renninger, 2006; Krapp, 2002; Mitchell, 1993; Schiefele, 1991). We agree, but believe these constructs can be distinguished over time. In particular, we are interested in how the perception of task value early in a learning experience influences the development of subsequent interest. For example, students might find a specific lecture or reading assignment in a psychology class to be enjoyable (intrinsic value) or relevant to their lives (utility value), and this perception of value could then lead to the development of subsequent interest in psychology. In terms of Hidi and Renninger's (2006) four-phase model of interest development, perceived task values would correspond to both the triggering of situational interest and the early stages of maintained situational interest, whereas subsequent interest in psychology would correspond to the latter stages of maintained situational interest or the emergence of individual interest. This conceptualization of task values as situational interest is not inconsistent with Eccles et al.'s (1983) perspective; rather, it is simply a more specific formulation that allows us to integrate these theoretical perspectives. In the Eccles et al. (1983) model of task values, intrinsic and utility task values could be considered as more general beliefs or as situation-specific reactions to the task. In this article, we consider task values to be situation-specific predictors of subsequent interest and performance; whereas *interest* refers to more general beliefs about the activity over time. This important distinction allows us to consider perceived task value and subsequent interest as separate outcomes, and to evaluate how the emergence of value during task engagement contributes to the development of subsequent interest.

This model of interest development is supported by our previous theoretical (Hidi & Harackiewicz, 2000) and empirical (Harackiewicz et al., 2008) work, and requires that we consider the initial level of interest that students and athletes bring into learning situations. Hidi and Renninger's (2006) four-phase model of interest development outlines how interest develops and deepens in a situation depending on the value, positive affect, and knowledge experienced with an activity. Individuals who have different levels of experience and knowledge of an activity may enter a situation with different levels of initial interest. Thus, in order for us to assess the role of value in interest development, we will need to consider initial levels of interest in the activity (Hidi & Renninger, 2006; Maxwell & Cole, 2007).

The integration of achievement goal, interest, and expectancy-value models of motivation complements the previously established notion that there can be multiple pathways to optimal motivation (Barron & Harackiewicz, 2001; Harackiewicz, Barron, Pintrich, et al., 2002; Pintrich, 2000b, 2003). For example, achievement goal researchers have emphasized the importance of testing the independent and interactive effects of achievement goals on outcomes, and have identified several patterns of multiple goal effects (cf. Harackiewicz, Barron, Pintrich, et al., 2002). In

this article, we also assess the indirect effects of achievement goals through task values. This analysis complements the multiple goals perspective in that it provides additional pathways through which achievement goals might promote interest and performance. Although not exhaustive, our theoretical integration incorporates three related perspectives in which we have conducted prior research, and as such is an initial step toward theoretical synthesis. In addition, our framework is generally consistent with other motivational models, such as those that propose dual-regulation systems (Boekaerts, 2003; Boekaerts & Corno, 2005; Heckhausen, 2000; Krapp, 2002, 2005; Kuhl, 2001; Schiefele, 1991). These models suggest that behavior is regulated by both cognitive/rational and emotional/feeling subsystems, which parallels our explication of utility and intrinsic value as two components of situational interest.

### Current Research

In two studies, we investigated the predictors and consequences of task value in classroom and sports contexts. We were primarily interested in (a) the role of task value in interest development and performance, and (b) the role of achievement goals in creating a framework for the perception of task value. In Study 1, we examined college students' perceptions of task values in an introductory psychology course, as well as their achievement goals, as predictors of their subsequent interest in the topic and academic performance. In Study 2, we examined high school athletes' perceptions of task values at a summer football camp, as well as their achievement goals, as predictors of their subsequent interest and performance at camp. In both studies, we controlled for initial interest in the activity.

College classes and sports camps are similar in terms of their educational mission and definition of achievement. In both settings, there is a focus on learning: Students learn lessons, do homework, and take tests, whereas athletes learn techniques, practice and do drill work, and play competitive games. For some individuals, sports may be more interesting than academics (and vice versa), but both can involve long hours of homework and repetitive drills. Success in both domains can require the ability to practice and work through difficult or boring activities. Given these similarities, theories of achievement motivation and interest should be applicable to both domains.

There is also an important difference between sports and school that should not be overlooked, particularly in terms of interest development: Most athletes have chosen to participate in their sport, whereas students often have little or no choice in which classes they take, particularly when fulfilling academic requirements. Although the decision to take an introductory course may indicate a certain level of initial interest in the topic, most students are unlikely to have had extensive experience with the topic. The introductory course therefore provides the opportunity to examine the emergence and development of interest in its early phases (Hidi & Renninger, 2006). In contrast, athletes willing to attend a summer camp at their own expense typically have extensive experience with the sport and a well-developed interest in it. Thus, the high school summer sports camp provides the opportunity to examine the role of situational interest in deepening already well-developed interest.

## Study 1—The College Classroom

We investigated the role of task values and achievement goals in predicting subsequent interest and performance in a college classroom, controlling for initial levels of interest. On the basis of previous findings (Harackiewicz, Barron, Pintrich, et al., 2002; Wigfield & Eccles, 2002), we hypothesized that mastery goals would predict the perception of value and the development of subsequent interest in the course, and that performance-approach goals would predict course grades. Including a measure of initial interest allowed us to clarify the relationship between mastery goals, task values, and the development of subsequent interest. In addition, we hoped to explore relationships between values and performance by examining utility and intrinsic values separately. Specifically, we hypothesized that perceiving utility value in the course might also predict performance.

### Method

#### Overview

This study took place during the course of a semester at a large, midwestern university and consisted of three waves of data collection during the semester. Students' initial interest in the course topic and achievement goals were assessed on the second day of class (Wave 1); their perceptions of course value (intrinsic and utility) were measured several weeks into the semester, but before the first exam (Wave 2); and their interest in the course was measured during the last week of class (Wave 3). We also obtained students' final course grades from department records.

#### Participants and Setting

Participants were recruited from four sections of introductory psychology classes (approximately 350 students per section). Only students who were taking the course for graded credit were included in the sample. Classes were almost entirely lecture format. Students' grades were determined by their performance on several multiple-choice exams given throughout the semester. Final grades were assigned based on a normative curve recommended by the psychology department.

Wave 1 data collection was part of a departmental-wide survey (1,145 participants) for which students received extra credit in exchange for participation. Participants enrolled in our study at Time 2 (830 participants) when they completed consent forms and the accompanying survey during class time. Students were not compensated for their participation in the second and third waves. Some attrition did occur after enrollment in our study at Wave 2. First, 35 participants dropped the course before the end of the semester. Second, because data were collected during class time, students who were absent at Wave 3 ( $n = 132$ ) were not included in the study. Thus, the final sample used in data analysis included 663 students (215 men and 448 women) and represents 79.9% of the students who enrolled in our study at Wave 2. Additionally, 2 participants who received a grade of "incomplete" were not included in the course grade analysis.

#### Measures

**Initial measures.** During the second meeting of the course, after students had become familiar with the course syllabus and

materials, students completed a questionnaire that contained items assessing their initial interest in psychology, and their achievement goals for the semester. Initial interest in psychology was assessed with two items ("I think psychology is a very interesting subject," "I don't think psychology is a very interesting subject [reversed]";  $\alpha = .78$ ). Mastery-approach goals were assessed with two items ("My goal is to learn as much as possible about psychology," "I want to develop an understanding of psychology";  $\alpha = .80$ ). Performance-approach goals were assessed with two items ("I would like to do better than other students in this class," "My goal is to get a better grade than other students in this class";  $\alpha = .75$ ). The initial interest and achievement goal items were based on prior research (Harackiewicz, Barron, Tauer, & Elliot, 2002; Harackiewicz et al., 2008). Participants responded to all self-report items on a scale from 1 (*strongly disagree*) to 7 (*strongly agree*).

**Midsemester measures.** Four weeks into the course, and about 1 week prior to the first exam, students completed a questionnaire that contained items assessing their perceived values in the course and their interest in the course. Utility value was assessed with three items ("What I am learning in this class is relevant to my life," "The topics in this class are important for my career," "In general, material from this class is not useful to me [reversed]";  $\alpha = .72$ ). Intrinsic value was assessed with three items ("Lectures in this class are entertaining," "Lectures in this class drag on forever [reversed]," "I enjoy coming to lecture";  $\alpha = .82$ ). Interest in the course was assessed with the same two items used for the initial measure of interest ( $\alpha = .78$ ).

**Subsequent interest.** At the end of the semester, but prior to the final exam, students' final interest in the course was assessed. In addition to the items used to measure initial interest, the measure of subsequent interest included two behavioral inclination items ("I would like to take more psychology courses," "My experience in this course has made me want to take more psychology courses";  $\alpha = .87$ ).

**Final grade.** Students' final course grades were obtained from departmental records. Each student could receive one of seven possible grades, based on the university's 4-point scale. The average grade in this study was 2.92 ( $SD = 0.84$ ). Grades were distributed as follows: A = 21.5%, AB = 19.5%, B = 16%, BC = 17.4%, C = 20.7%, D = 4.7%, and F = 0.2%. Two students in our sample received an incomplete grade for the semester, but were retained in the interest analysis. The average grade for students in our sample was slightly higher than for all students in introductory psychology ( $M = 2.70$ ,  $SD = 0.97$ ,  $N = 1,267$ ), but the distribution of grades was similar.

### Results

#### Preliminary Analysis

**Attrition.** We conducted an analysis that compared the final sample ( $n = 663$ ) with those individuals who enrolled in our study at Time 2 but who were not included in the final sample because they either dropped the course or failed to attend class on the day of data collection ( $n = 167$ ). Comparisons were made on all the major variables in the study: initial interest, mastery-approach and performance-approach goals, intrinsic and utility value, subsequent interest, and final grade.

Independent-sample *t* tests revealed two significant differences. Effect sizes (Cohen's *d*) were computed for these differences,



revealing small to large effects. The final sample had higher scores on mastery-approach goals,  $t(825) = -2.57, p = .01, d = .24$ , and course grade,  $t(793) = -7.28, p < .01, d = .82$ .

*Descriptive and correlational analyses.* Means, standard deviations, correlations, and scale reliabilities for the variables in this study are presented in Table 1. Although the structure and content and grading distributions of the four classes were comparable, we tested for instructor differences on all variables. There were significant instructor effects on intrinsic value, utility value, and subsequent interest, indicating that students reported higher levels of intrinsic value, utility value, and subsequent interest with some instructors compared with others. Therefore, we included a set of three dummy-code terms to test and control for mean-level differences between the instructors in all subsequent analyses (Cohen, Cohen, West, & Aiken, 2003).

*Missing data.* It is possible that the attrition that occurred in this data set could be influencing the findings in a myriad of ways. A complete account of causality and interpretation issues due to missing data can be found elsewhere (Rubin, 1976). The number of people with available data is presented for each variable in Table 1. We created a dichotomous variable reflecting whether individuals had complete data across all study variables (coded 0) or had incomplete data on at least one variable (coded 1). This variable is included in Table 1. From the correlations between this variable and the other variables we can determine the extent to which incomplete data are related to other variables in the study. The bottom row of correlations shows that students with a lower final grade, lower initial interest, and lower mastery goals were more likely to have incomplete data. These associations indicate that the incomplete data are not missing completely at random, because if they were, then the missing pattern would not be related to any variables in the data set (Rubin, 1976). In addition, these associations enabled us to create a new data set that accounted for at least some of the variability in the missing data.

One approach to missing or incomplete data was to create imputed data sets that model what the data may have looked like if none of the data were missing. We used Mplus (Version 3.01), using full information maximum likelihood, to impute a complete data set that includes values on each variable for each individual

(Muthén & Muthén, 2006). The results of that analysis did not diverge from the analysis of only the originally complete data. Therefore, in this article, we report only the analyses using the original data set that did not account for missing data. A complete report of the maximum likelihood analysis is available from Chris S. Hulleman on request.

*Scale construction.* To test our hypothesis that the interest, utility value, and intrinsic value items would form separate factors, we conducted two types of factor analyses. First, in order to examine the factor loadings on our hypothesized three-factor model, we conducted an exploratory factor analysis of the items used to measure subsequent interest, utility value, and intrinsic value using a separate sample of undergraduates. As a part of a larger study on student attitudes toward learning, participants ( $N = 188$ ) were recruited from four sections of an introductory statistics course, and completed a brief survey during class time. The items and the pattern matrix of factor loadings are presented in Table 2. Each item loaded onto its respective factor (all values  $\geq .678$ ), and there was an absence of cross-loadings onto other factors (all cross-loadings  $< .300$ ).

Second, in order to test how well our model fit the data, we conducted a confirmatory factor analysis (CFA) using LISREL 8.72 (Jöreskog & Sörbom, 1998) on the current sample of introductory to psychology students ( $N = 663$ ). Five nested models were tested. The three-factor model—which separated interest, intrinsic value, and utility value—was tested against all other possible factor groupings: (a) a single factor that combined interest, intrinsic, and utility value items; (b) two factors: one interest factor and one combined intrinsic/utility value factor; (c) two factors: one utility value factor and one combined interest/intrinsic value factor; and (d) two factors: one intrinsic value factor and one combined interest/utility value factor. For each model, we calculated multiple indices of fit: chi-square, comparative fit index, standardized root-mean-square residual, and root-mean-square error of approximation. The results of the five confirmatory factor analyses are presented in Table 3. The hypothesized model—with distinct interest, utility, and intrinsic value factors—provided a better fit than any of the other models. The worst fitting model was the one that combined all three factors into one combined factor.

Table 1  
Means, Standard Deviations, and Correlations for Major Variables in Study 1

Variable	1	2	3	4	5	6	7	8
1. Initial interest	<i>0.78</i>							
2. Mastery-approach goals	0.62	<i>0.80</i>						
3. Performance-approach goals	0.09	0.21	<i>0.75</i>					
4. Utility value	0.36	0.38	0.07	<i>0.72</i>				
5. Intrinsic value	0.21	0.20	0.06	0.42	<i>0.82</i>			
6. Final grade	0.04	0.04	0.17	0.17	0.06	—		
7. Subsequent interest	0.44	0.34	0.07	0.47	0.33	0.29	<i>0.87</i>	
8. Missing data code	−0.08	−0.09	−0.01	−0.04	0.09	−0.24	−0.02	—
Minimum	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.00
Maximum	7.00	7.00	7.00	7.00	7.00	4.00	7.00	1.00
<i>M</i>	6.00	5.99	5.80	5.11	4.86	2.92	5.18	0.57
<i>SD</i>	1.10	1.02	1.25	1.04	1.14	0.84	1.25	0.49
<i>n</i>	1,150	1,145	1,145	830	830	661	663	1,150

Note. Correlations .08 and larger are significant at  $p < .01$ . Values along the main diagonal (in italics) are scale reliabilities. Missing data code: 0 = complete data, 1 = incomplete data. Dashes indicate that it was not possible to compute reliabilities.

Table 2  
Factor Analysis Pattern Matrix Factor Loadings from Statistics Sample (Study 1)

Variable	Factor		
	Interest	Utility	Intrinsic
Interest 1	.964		
Interest 2	.879		
Interest 3	.787		
Interest 4	-.678		
Utility 1		.843	
Utility 2		.728	
Utility 3		.696	
Utility 4		.689	
Intrinsic 1			.873
Intrinsic 2			.780
Intrinsic 3			-.699
Factor correlations			
Utility value	0.57		
Intrinsic value	0.59	0.42	

*Note.* Values are pattern matrix factor loadings from a factor analysis using principal-axis factoring with oblique rotation. Values less than .300 were omitted from this table. Interest 1 = "To be honest, I just don't find statistics interesting"; Interest 2 = "I think what we're learning in this class is fascinating"; Interest 3 = "I think the field of statistics is very interesting"; Interest 4 = "Statistics fascinates me"; Utility 1 = "I can apply what we are learning in Introductory Statistics to some of my other courses"; Utility 2 = "What I am learning in this class is relevant to my life"; Utility 3 = "I can apply what we are learning in Introductory Statistics to real life"; Utility 4 = "I think what we are studying in Introductory Statistics is useful for me to know"; Intrinsic 1 = "Lectures in this class are entertaining"; Intrinsic 2 = "I enjoy coming to lecture"; Intrinsic 3 = "Lectures in this class drag on forever."

### Regression Analyses

We used hierarchical multiple regression to analyze these data in three stages. First, we investigated the antecedents of achievement goals. Second, we examined the direct effects of task values, achievement goals, and initial interest on subsequent interest and course grade. Finally, we tested the indirect effects of all variables on subsequent interest and course grade. For all analyses, all continuous variables were standardized, and all two- and three-way interaction terms were computed (Aiken & West, 1991).

Significant interaction effects were interpreted in two ways. First, we computed predicted values for representative high and low groups (one standard deviation above and below the mean) from the regression equations using the unstandardized regression coefficients. Second, we calculated simple slope values for representative high and low groups from the regression equations and tested them according to Aiken and West (1991).

*Predictors of achievement goals.* The first set of analyses examined the antecedents of achievement goals. The achievement goal model contained the three instructor dummy codes, initial interest, and an effects code for gender (men = -1, women = +1). This five-term achievement goal model was used to predict both mastery- and performance-approach goals. Initial interest was the only significant predictor for both mastery-approach,  $t(662) = 19.69$ ,  $p < .01$  ( $\beta = .61$ ), and performance-approach,  $t(662) = 2.64$ ,  $p = .01$  ( $\beta = .11$ ) goals. Individuals with higher levels of initial interest reported higher levels of mastery- and performance-approach goals.

*Direct effects models.* The second set of analyses tested the direct effects of initial interest, mastery- and performance-approach goals, and all of their two- and three-way interactions on subsequent interest and final grade. We also included the gender effects code and the three instructor dummy codes. Preliminary testing indicated that none of the three-way interactions between initial interest, achievement goals, and gender attained significance on any measure, and they were dropped from all subsequent analyses. All two-way interactions were retained in the models. Therefore, the direct effects model contained 10 terms: the 7 main effect terms as well as 3 two-way interactions (Mastery-Approach  $\times$  Performance-Approach, Mastery-Approach  $\times$  Initial Interest, Performance-Approach  $\times$  Initial Interest).

*Direct effects on subsequent interest.* There was a significant main effect of initial interest,  $t(652) = 9.86$ ,  $p < .01$  ( $\beta = .46$ ), such that students who entered the course with higher levels of initial interest reported more interest in psychology at the end of the semester. The main effect of mastery-approach goals,  $t(652) = 3.60$ ,  $p < .01$  ( $\beta = .17$ ), indicated that students with higher levels of mastery goals at the beginning of the course were more interested in psychology. These two main effects were qualified by the significant interaction between mastery-approach goals and initial interest,  $t(652) = 5.35$ ,  $p < .01$  ( $\beta = .25$ ), which indicated that mastery-approach goals were a stronger predictor of subsequent

Table 3  
Confirmatory Factor Analysis Results from Study 1

Model	$\chi^2$ (df)	$\Delta\chi^2$ (df) <sup>a</sup>	RMSEA	CFI	SRMR
One factor	601.54** (20)	522.70** (3)	0.21	0.85	0.15
Two factor: Interest and value	472.84** (19)	394.00** (2)	0.19	0.89	0.14
Two factor: Utility value and combined interest/intrinsic	443.32** (19)	364.48** (2)	0.19	0.86	0.13
Two factor: Intrinsic value and combined interest/utility	168.41** (19)	89.57** (2)	0.11	0.96	0.08
Three factor	78.84** (17)		0.08	0.98	0.06

*Note.*  $N = 663$ . RMSEA = root-mean-square error approximation; CFI = comparative fit index; SRMR = standardized root-mean-square residual.

<sup>a</sup> Test of the difference in model fit with the three-factor model (separate interest, utility value, and intrinsic value factors).

\*\* $p < .01$ .



interest in psychology for those students with high levels of initial interest ( $\beta = .42$ ) than for those with low levels of initial interest ( $\beta = .01$ ). This interaction is graphed in the top panel of Figure 1.

**Direct effects on final grade.** There was a significant main effect of performance-approach goals,  $t(650) = 4.56, p < .01$  ( $\beta = .18$ ), such that students with higher levels of performance-approach goals obtained higher final course grades. The direct effects path model for final interest and final grade is presented in Figure 2. Although we tested for the effects of initial interest in all analyses, we did not include it in the path diagrams to enhance parsimony.

**Mediated (and indirect effects) models.** The third set of analyses tested the role of perceived task values as mediators or indirect paths between initial interest and goals and the final outcome measures. First, we tested whether achievement goals and initial interest were significantly related to task values by using the direct effects regression model to predict intrinsic and utility value. Second, we tested whether the mediators were related to the outcomes, while controlling for the original predictors, by including task values in the direct effects model. Therefore, we added intrinsic and utility value to the direct effects model, as well as all possible two- and three-way interactions between values, initial interest, and goals. For each set of analyses, only significant interactions were retained in the model. Thus, unless otherwise noted, the mediation model contained 12 terms: 10 terms from the direct effects model with the addition of the 2 value terms. Finally, to test the significance of mediated and indirect effects, we followed procedures outlined by Kenny, Kashy, and Bolger (1998) in which the indirect path is tested by creating a product term and dividing by its standard error. This technique has been shown to be robust to Type I errors (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002).

**Direct effects on intrinsic value.** The significant main effects of initial interest,  $t(652) = 3.13, p < .01$  ( $\beta = .16$ ), and mastery-approach goals,  $t(652) = 2.53, p = .01$  ( $\beta = .13$ ), indicated that students with higher levels of initial interest and mastery-approach goals were more likely to find intrinsic value in the course topic. These main effects were qualified by their significant interaction,  $t(652) = 2.08, p = .038$  ( $\beta = .11$ ), which indicated that mastery-approach goals led to more intrinsic value for those students with high levels of initial interest ( $\beta = .24$ ), but not for those with low levels of initial interest ( $\beta = .02$ ). This interaction is presented in the middle panel of Figure 1.

**Direct effects on utility value.** The significant main effects of initial interest,  $t(652) = 6.32, p < .01$  ( $\beta = .30$ ), and mastery-approach goals,  $t(652) = 6.68, p < .01$  ( $\beta = .32$ ), indicated that students with higher levels of initial interest and mastery-approach goals were more likely to find utility value in the course topic. These main effects were qualified by their significant interaction,  $t(652) = 5.86, p < .01$  ( $\beta = .28$ ), which indicated that initial interest led to more utility value for those students with high levels of initial interest ( $\beta = .60$ ), but not for those with low levels of initial interest ( $\beta = .04$ ). This interaction is presented in the lower panel of Figure 1.

**Mediated and indirect effects on subsequent interest.** In the mediation model, the significant main effects of intrinsic value,  $t(650) = 3.84, p < .01$  ( $\beta = .14$ ), and utility value,  $t(650) = 6.33, p < .01$  ( $\beta = .25$ ), indicated that students who perceived each type of value in the course reported more interest in psychology at the

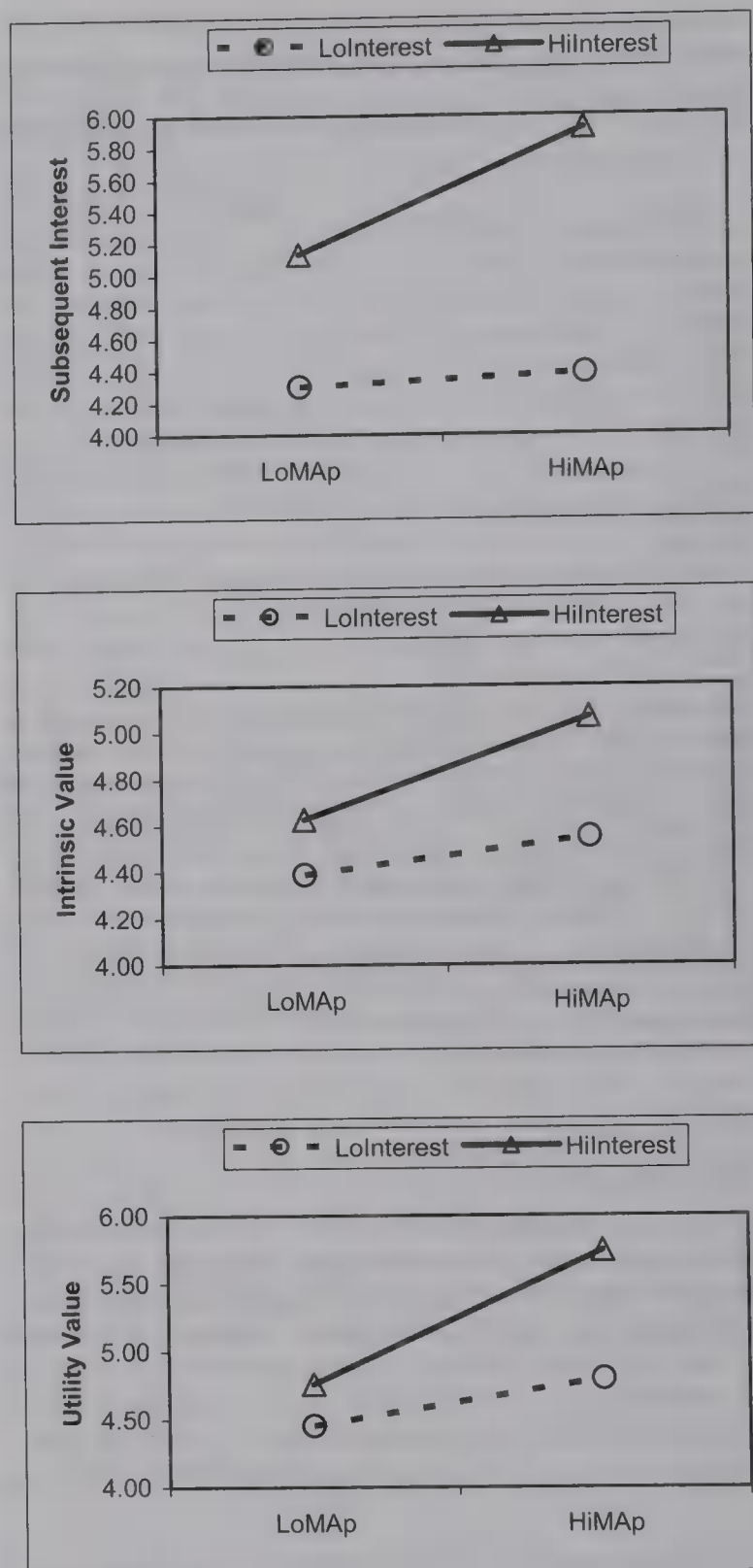


Figure 1. Interactive effects of mastery-approach goals and initial interest on subsequent interest and task values in Study 1. Interest = interest in psychology at the beginning of the semester; MAp = mastery-approach goals; PAp = performance-approach goals. The *Hi* and *Lo* prefixes indicate values at one standard deviation above and below the mean, respectively, for each variable. Each panel represents a significant two-way interaction between mastery-approach goals and initial interest. The top panel represents the interaction on subsequent interest, the middle panel represents the interaction on intrinsic value, and the bottom panel represents the interaction on utility value. The simple slopes for initial interest at one standard deviation above the mean of mastery-approach goals ( $\beta = .61, .28, .58$ ) and at one standard deviation below the mean of mastery-approach goals ( $\beta = .21, .05, .02$ ) were calculated for subsequent interest, intrinsic value, and utility value, respectively.

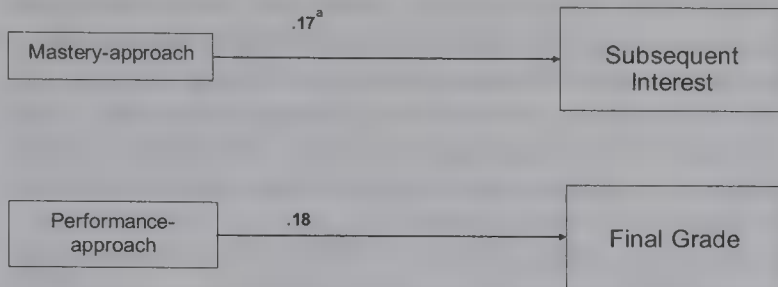


Figure 2. Study 1 path model of direct effects on course outcomes. <sup>a</sup>The significant mastery-approach direct effect on subsequent interest was qualified by its significant interaction with initial interest (see Figure 1 and the text). All paths are standardized regression coefficients from a simultaneous multiple regression that includes all direct effects in the model. All paths are significant at  $p < .01$ .

end of the semester. The main effect of initial interest was also significant,  $t(650) = 8.03$ ,  $p < .01$  ( $\beta = .36$ ), but reduced in size from the direct effects model ( $\beta = .46$ ). The formal test of mediation revealed that the effect of interest was partially mediated by both intrinsic value ( $z = 2.43$ ,  $p = .02$ ) and utility value ( $z = 4.47$ ,  $p < .01$ ). In addition, the mastery-approach goal main effect was no longer significant ( $\beta = .07$ ,  $p = .12$ ) and reduced in size from the direct effects model ( $\beta = .17$ ). The formal test of mediation revealed that the effect of mastery goals was partially mediated through both intrinsic ( $z = 2.11$ ,  $p = .03$ ) and utility value ( $z = 4.49$ ,  $p < .01$ ). The complete path model is presented in Figure 3. The interaction between initial interest and mastery goals remained significant in the mediated model,  $t(650) = 3.68$ ,  $p < .01$  ( $\beta = .17$ ), but reduced in size from the direct effects model ( $\beta = .25$ ). The formal test of mediation revealed that this interaction effect was partially mediated through both intrinsic value ( $z = 1.83$ ,  $p = .07$ ) and utility value ( $z = 4.30$ ,  $p < .01$ ).

**Mediated and indirect effects on final grade.** In the mediated model, the significant main effect of performance-approach goals was unchanged from the direct effects model,  $t(648) = 4.44$ ,  $p < .01$  ( $\beta = .18$ ). However, there was a significant main effect of utility value,  $t(648) = 3.94$ ,  $p < .01$  ( $\beta = .19$ ), such that students who perceived higher levels of utility value in psychology received

higher grades than students who perceived lower levels of utility value. Although initial interest and mastery-approach goals did not directly predict final grade, there were significant indirect paths from these variables to final grade through utility value ( $z = 3.34$ ,  $p < .01$ , and  $z = 3.39$ ,  $p < .01$ , respectively). The complete path model for final interest and final grade is presented in Figure 3.

### Structural Equation Model

In order to test the overall fit of the model, we used Lisrel 8.72 to simultaneously test the model presented in Figure 3. The results indicated that the overall model provided a satisfactory fit to the data,  $\chi^2(80, N = 663) = 309.44$ ,  $p < .001$ , root-mean-square error of approximation = .07, comparative fit index = .97, standardized root-mean-square residual = .07.

### Discussion—Study 1

The results of this study suggest that perceptions of task value play an important role in the development of subsequent interest, and are also associated with academic performance. Adopting mastery-approach goals was associated with the perception of utility and intrinsic value in the course, particularly for students already high in initial interest, and these task values were in turn associated with the development of subsequent interest. Perceptions of utility value were also positively associated with final grades. Taken together, these results highlight the important role of mastery goals in optimal motivation as they promote the perception of value.

This study replicated prior work on achievement goal effects in the college classroom and demonstrated that mastery-approach goals predicted subsequent interest in the course, whereas performance-approach goals predicted higher course grades (Harackiewicz, Barron, Pintrich, et al., 2002). In addition, this study extends prior research in several important ways. First, when we controlled for initial interest and achievement goals, utility and intrinsic task values were unique predictors of subsequent interest. In addition, the effects of both initial interest and mastery-approach goals on subsequent interest were mediated through the value students perceived in the course. Initial interest and mastery-

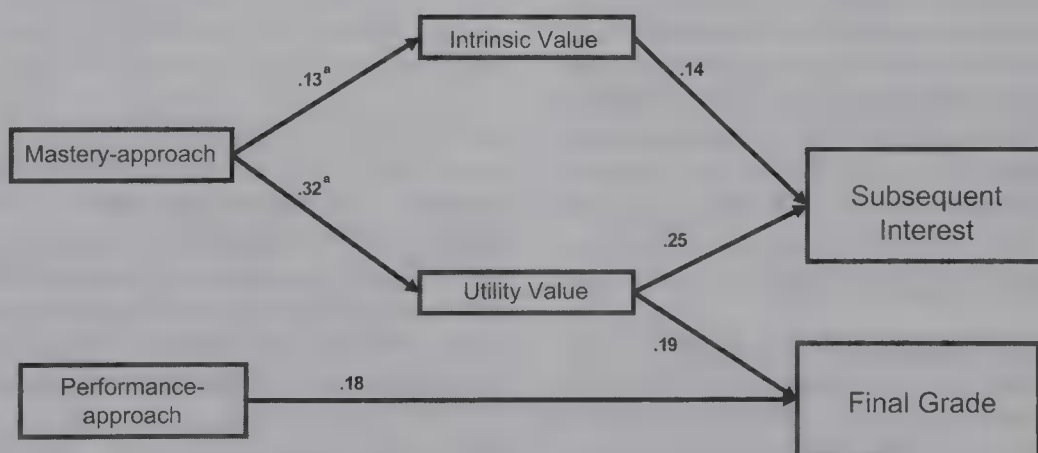


Figure 3. Study 1 path model of direct and mediated effects on course outcomes. <sup>a</sup>The direct effects of mastery-approach goals on intrinsic and utility value were moderated by significant interactions with initial interest (see text for details). All paths are standardized regression coefficients from simultaneous multiple regressions that included all the variables in the model. All paths are significant at  $p < .01$ .



approach goals predicted utility and intrinsic value in the course material, which predicted subsequent interest, supporting the four-phase model of interest development (Hidi & Renninger, 2006).

Second, perceiving utility value in the course was positively associated with higher course grades. This finding represents an indirect pathway through which initial interest and mastery-approach goals positively influenced academic performance. It is important to note that our results are not informative about the processes whereby task values were associated with subsequent interest and performance. Perhaps finding the material useful led students to make more meaningful connections between themselves and the material, which allowed them to learn the material more thoroughly, remember it better (Markus, 1977), and put more effort into the course (Mac Iver et al., 1991). Future research will need to explicate the mechanisms of the task value effects.

Third, the effects of mastery-approach goals on subsequent interest remained significant even when we controlled for initial interest. In fact, despite being highly correlated ( $r = .61, p < .01$ ), initial interest and mastery-approach goals had unique, independent effects on subsequent interest. In addition, the interaction between initial interest and mastery-approach goals revealed that mastery-approach goals were strongly predictive of subsequent interest only for students who were high in initial interest. In contrast, mastery-approach goals were unrelated to subsequent interest for individuals who entered the class with low levels of initial interest. Thus, the combination of initial interest and mastery-approach goals was particularly powerful in predicting task values and subsequent interest.

In Study 2, we sought to extend our analyses to a different achievement context, and we investigated task values and achievement goals at a sports camp. Participants were high school football players who participated in a summer football camp designed to prepare them for the upcoming season. In contrast to introductory psychology students in Study 1, the football players in Study 2 were more likely to have extensive experience playing football and thus entered the camp situation with more well-developed interest. This allowed us to investigate a proposition of Hidi and Renninger's (2006) model regarding the deepening of already well-developed individual interest. According to their model, well-developed individual interest is maintained and deepened through continued interaction with the activity: By re-engaging in the activity, individuals with well-developed interests can experience new-found situational interest—through knowledge and skill acquisition, finding new meaning and value, and experiencing positive affect—and thereby deepen their interest in the activity. In Study 2, we investigated the first step in this process by examining the development of situational interest in a domain in which individuals begin with a more well-developed individual interest.

### Study 2—Sports Camp

This study investigated the role of task values and achievement goals in predicting subsequent interest and performance at a summer football camp for high school boys, controlling for initial interest. High school sports camps provide opportunities for athletes to engage in their sport outside of the competitive season and to develop and extend their existing skills. These camps tend to attract individuals with high levels of experience and interest in the sport. Indeed, not only did our athletes report being interested in

football ( $M = 5.54, SD = 0.78$ ), but they also had been playing the sport for several years ( $M = 4.26, SD = 1.93$ ). Those campers who get the most out of camp are likely to leave camp with new skills as well as with a sense of enjoyment and satisfaction, having optimized their camp experience. Thus, our primary outcome measure in this study was camp satisfaction, reflecting situational interest in the camp experience.

This study extends prior achievement goal research by examining the influence of both types of approach achievement goals in the sports context. Prior research in sports contexts has assessed mastery-approach goals, but performance goals have typically been assessed as a combination of approach and avoidance motivation (Elliot, 2005) or within the ego-orientation perspective (Duda, 1995). On the basis of prior research in the classroom, we predicted that mastery-approach goals would predict satisfaction in a sports camp, whereas performance-approach goals would predict performance at camp. We also hypothesized that mastery goals would predict both intrinsic and utility task values. Consistent with the results of Study 1, we hypothesized that intrinsic value would predict camp satisfaction and that utility value would predict both camp satisfaction and performance.

In addition, because this research extends beyond the classroom, we can speculate about how achievement goal effects might differ in this particular sports context. Sporting contexts can differ from classrooms in that competition, rewards, and performance are inherent to the activity (Hidi, 2000; Kruglanski, 1975), and these components of the activity might enhance subsequent interest and intrinsic motivation (Hidi & Harackiewicz, 2000). Therefore, it is possible that performance-approach goals might positively predict camp satisfaction. However, it is also possible that mastery-approach goals might predict performance in this context. This is because football camp occurs outside of the regular competitive season and skill development could be more salient than normative performance at football camp, and more salient than in the classroom. Thus, this particular sporting context, in which skill development is an inherent part of the purpose of the activity, could provide the opportunity for achievement goals to work in ways that are different from the college classroom and reveal multiple pathways for optimal motivation (Barron & Harackiewicz, 2001; Hidi & Harackiewicz, 2000; Pintrich, 2000b).

## Method

### Overview

This study took place during a summer football camp for high school boys in the midwestern United States. Participants' initial interest in football and achievement goals were assessed on the first day of the camp (Wave 1), their task values for specific football drills were assessed during the middle of the camp (Wave 2), and their interest in camp was assessed at the conclusion of camp (Wave 3). Our dependent measure of performance consisted of coaches' ratings of campers' effort and performance at camp.

### Participants and Setting

Participants were recruited from two summer football camps for high school boys (approximately 200 campers per camp). Participants were entering 8th–12th grade in the coming fall (grade level



of participants: 8th = 0.8%, 9th = 14.3%, 10th = 24.4%, 11th = 26.9%, 12th = 33.6%). The football camps were either 3- or 4-day overnight camps that focused on teaching players football-relevant skills. Each day of camp included two or three practice sessions that were focused on skill development, allowing only minimal opportunity to scrimmage or play competitive games. Camp drills were divided into two categories: technique drills and competitive drills. Technique drills were defined as structured practice drills in which players were working individually on form or footwork, and not competing against someone. Examples of technique drills are form running, footwork drills, running routes, or pushing a blocking sled. Competitive drills were defined as structured practice drills in which players were competing against others, but not actually scrimmaging as an entire team. Examples of competitive drills are one-on-one passing and blocking drills.

Participants enrolled in our study when they completed the Wave 1 consent form and questionnaire on the first day of camp (237 participants). Players were not compensated for their participation in any of the surveys. Of the original 237 participants, 198 completed the Wave 2 measures, and 155 completed the Wave 3 measures. The participant attrition was primarily due to players needing to leave camp early because of schedule conflicts or participation in other summer sports events. Because the coach ratings were not dependent on participant attrition, the final sample for coach ratings was higher ( $n = 198$ , 84% of the original sample) than that for camp interest ( $n = 155$ , 65% of the original sample).

## Measures

*Initial interest and achievement goals.* Participants were mailed an initial questionnaire to complete and return when they checked into camp. This measure assessed participants' interest in football as well as mastery- and performance-approach goals for football camp. We adapted the initial interest measures for use in the football context from the final interest measures used in Harackiewicz, Barron, Tauer, and Elliot (2002), for example, "Football is my favorite sport to play." The specific items in this six-item scale had good reliability ( $\alpha = .78$ ) and can be found in the Appendix. Mastery-approach goal items focused on skill development at camp ("I want to develop my football skills at this camp," "I want to learn as much as possible at this camp," "I don't really care how much I learn at this camp [reversed]";  $\alpha = .72$ ). Performance-approach goal items focused on normative comparisons at camp ("It is important for me to do well compared to others at this camp," "My goal is to be one of the best football players at this camp," "I would like to do better than other players at this camp," "I don't care how I do compared to other players at this camp [reversed]";  $\alpha = .79$ ). The achievement goal items were adapted for the camp context based on Harackiewicz, Barron, Tauer, and Elliot (2002) and Conroy, Elliot, and Hofer (2003). All self-report measures in this study used a 7-point Likert scale that ranged from 1 (*strongly disagree*) to 7 (*strongly agree*).

*Utility and intrinsic value.* Approximately halfway through the camp participants completed a 10-item questionnaire that contained the task value measures. These value items were based on the model of values proposed by Eccles and Harold (1991) and assessed participants self-reported utility and intrinsic values of specific drills at camp. The six-item utility value scale focused on how useful the drills were in attaining other ends, ("Technique/

competition drills are useful for my football career," "Technique/competition drills are a good opportunity to improve and develop my football skills," "Technique/competition drills are valuable in developing football skills";  $\alpha = .92$ ). The four-item intrinsic value scale focused on how enjoyable campers perceived the drills to be ("Technique/competition drills are fun," "Technique/competition drills are my favorite part of camp";  $\alpha = .72$ ).

*Camp satisfaction.* At the conclusion of camp, participants' enjoyment and satisfaction with their camp experience was assessed with five items (e.g., "I had fun at camp," "Overall, I am satisfied with this camp";  $\alpha = .77$ ). The specific items are presented in the Appendix.

*Coach ratings.* Coaches' evaluations of each participant's effort and performance over the course of the camp were measured at the conclusion of camp. Coaches identified the campers with whom they worked most often. Each player was rated on how much effort each put forth during camp, and how well he performed compared with other campers, on a scale from 1 (*little effort/low performance*) to 9 (*high effort/performance*). The ratings for effort and performance were  $z$  scored within coach and combined to form an overall composite rating of performance and effort for each player ( $\alpha = .77$ ).

## Results

### Preliminary Analysis

*Attrition.* We conducted an analysis that compared the final sample ( $n = 155$ ) with those individuals who were not included in the sample because they did not complete Wave 2 or 3 ( $n = 82$ ). Comparisons were made on all the major variables in the study: initial interest, mastery-approach and performance-approach goals, intrinsic and utility value, camp satisfaction, and performance. Independent-sample  $t$  tests revealed no significant differences on any variables.

*Descriptive and correlational analyses.* Means, standard deviations, correlations, and scale reliabilities for the variables in Study 2 are presented in Table 4. Although the purpose and structure of the camps was similar, we tested for camp differences on all variables. Independent-sample  $t$  tests revealed that one camp was significantly higher in initial interest than the other camp,  $t(235) = 4.12$ ,  $p < .01$  ( $d = .52$ ). Therefore, we included a camp effects code ( $-1, 1$ ) to test and control for mean-level differences between the two camps in all subsequent analyses (Cohen et al., 2003).

*Missing data.* We addressed the issue of missing data in Study 2 as we did in Study 1. The results of analyses using missing data techniques that included all participants (regardless of whether their data were complete) did not diverge from those that only included participants with complete data. In this article, we report only the analyses using the original data set that did not account for missing data.

*Factor analyses.* We conducted a confirmatory factor analysis using Lisrel 8.72 in order to test whether the intrinsic and utility value items separated into distinct components. Since we did not measure interest at the same time as value, we only included the task value measures in this analysis. We tested two competing models. The one-factor model specified that all 10 items loaded onto one overall value factor. The two-factor model specified that there were two independent value constructs representing utility



Table 4  
Means, Standard Deviations, and Correlations for Major Variables in Study 2

Variable	1	2	3	4	5	6	7	8
1. Initial interest	<i>0.75</i>							
2. Mastery-approach goals	0.20	<i>0.72</i>						
3. Performance-approach goals	0.26	0.00	<i>0.79</i>					
4. Utility value	0.28	0.23	0.17	<i>0.92</i>				
5. Intrinsic value	0.38	0.17	0.18	0.49	<i>0.72</i>			
6. Coach ratings	0.12	0.04	0.19	0.06	0.13	<i>0.77</i>		
7. Camp satisfaction	0.30	0.26	0.12	0.35	0.32	0.07	<i>0.78</i>	
8. Missing data code	-0.12	0.01	-0.08	-0.05	0.03	0.10	-0.14	—
Minimum	3.00	3.00	1.00	1.00	1.75	-2.63	3.20	0.00
Maximum	7.00	7.00	7.00	7.00	7.00	2.21	7.00	1.00
<i>M</i>	5.54	5.04	5.58	6.42	5.10	-0.02	6.33	0.39
<i>SD</i>	0.78	0.43	1.18	0.89	1.01	0.83	0.75	0.49
<i>N</i>	237	237	237	198	198	185	155	237

Note. Correlations .14 and larger are significant at  $p < .02$ . Values along the main diagonal (in italics) are scale reliabilities. Missing data code: 0 = complete data, 1 = incomplete data. The dash indicates that it was not possible to compute reliabilities.

and intrinsic value. As indicated in Table 5, neither model fit the data extremely well. However, the two factor model produced a significant better fitting model,  $\Delta\chi^2(1, N = 663) = 39.99, p < .01$ .

### Regression Models

We used hierarchical multiple regression to analyze these data. The primary analyses are divided into three parts. First, we investigated the antecedents of achievement goals. Second, we conducted a series of multiple regression analyses to examine the direct effects of initial interest, achievement goals, intrinsic value, and utility value on camp satisfaction and performance. Finally, we examined the indirect effects of all variables on camp satisfaction and performance. For all analyses, all continuous variables were standardized, and all two- and three-way interaction terms were computed (Aiken & West, 1991). Significant interactions were interpreted as in Study 1. Preliminary testing indicated that multicollinearity was an issue in this data set, in part because of the high degree of relationship between the value terms. For example, when all of the three-way interactions were entered simultaneously into the regression equation, tolerance values dropped to near zero for over half of the variables in the regression, which is an indicator of severe multicollinearity (Cohen et al., 2003). Therefore, when testing interactions, we were mindful of this problem and proceeded by testing the two-ways in one step, and the three-ways individually on a final step.

Table 5  
Confirmatory Factor Analysis Results from Study 2

Model	$\chi^2$ (df)	$\Delta\chi^2$ (df) <sup>a</sup>	RMSEA	CFI	SRMR
One factor	309.15** (35)	39.99** (1)	0.24	0.85	0.20
Two factor	269.16** (34)		0.23	0.87	0.18

Note.  $N = 155$ . RMSEA = root-mean-square error of approximation; CFI = comparative fit index; SRMR = standardized root-mean-square residual.

<sup>a</sup> Test of the difference in model fit with the two-factor model (separate utility and intrinsic value factors).

\*\*  $p < .01$ .

*Predictors of achievement goals.* The achievement goal model contained the camp effects code and initial interest. Initial interest was the only significant predictor for both mastery-approach,  $t(226) = 3.37, p = .001$  ( $\beta = .24$ ), and performance-approach,  $t(226) = 3.90, p < .01$  ( $\beta = .26$ ), goals. Individuals with higher levels of initial interest reported higher levels of mastery- and performance-approach goals, compared with those who had lower levels of initial interest.

*Direct effects models.* The second set of analyses tested the direct effects from initial interest, mastery- and performance-approach goals, and their two- and three-way interactions on camp satisfaction and performance. Nonsignificant interactions were trimmed from the final direct effects models. The camp effects code was also included.

*Direct effects on camp satisfaction.* There were significant main effects of initial interest,  $t(154) = 2.75, p < .01$  ( $\beta = .20$ ), and mastery-approach goals,  $t(154) = 3.10, p < .01$  ( $\beta = .22$ ). Players with higher levels of initial interest and mastery-approach goals reported more satisfaction with camp than those with lower levels of these variables.

In addition, there was a significant three-way interaction on camp satisfaction between initial interest, mastery-approach goals, and performance-approach goals,  $t(154) = -4.99, p < .01$  ( $\beta = -.38$ ). This interaction is graphed in Figure 4. Inspection of the predicted values indicated that players with low levels of initial interest reported the most camp satisfaction when both mastery and performance goals were high ( $\hat{Y} = 6.83$ ), compared with when both goals were low ( $\hat{Y} = 5.90$ ), when mastery goals were low and performance goals high ( $\hat{Y} = 5.93$ ), or when mastery goals were high and performance goals were low ( $\hat{Y} = 6.13$ ). In contrast, players with higher levels of initial interest demonstrated more camp satisfaction when only one goal type was high (high mastery/low performance,  $\hat{Y} = 6.83$ ; low mastery/high performance,  $\hat{Y} = 6.62$ ), and moderate levels of interest when both goals were high ( $\hat{Y} = 6.56$ ) or low ( $\hat{Y} = 6.43$ ).

*Direct effects on coach ratings.* There was a significant main effect of performance-approach goals,  $t(181) = 2.25, p = .03$  ( $\beta = .17$ ), such that players with higher levels of performance-approach goals received higher ratings of effort and performance from their

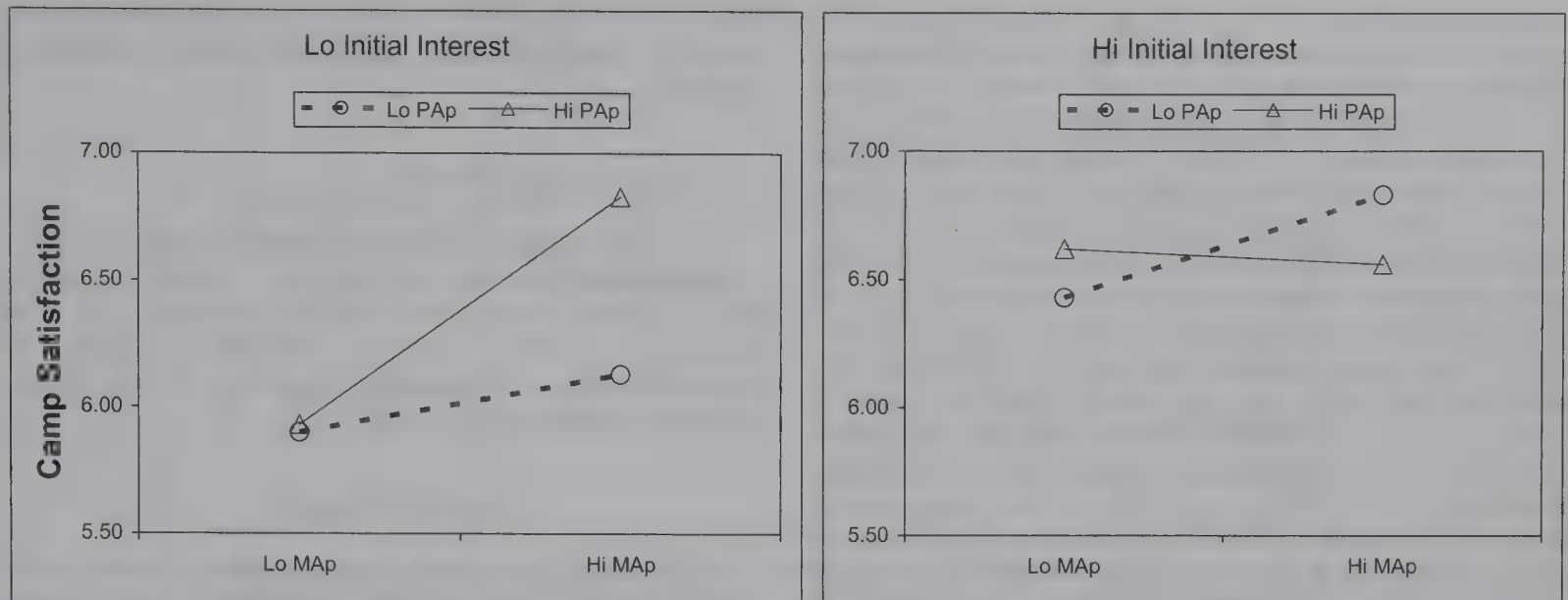


Figure 4. Interactive effects of initial interest, performance-approach, and mastery-approach goals on camp satisfaction in Study 2. The *Hi* and *Lo* prefixes indicate values at one standard deviation above and below the mean, respectively, for initial interest in football, mastery-approach goals (MAP), and performance-approach goals (PAp).

coaches. No other effects emerged. The direct effects path model for final interest and coach ratings is presented in Figure 5. Although we tested for the effects of initial interest in all analyses, we did not include it in the path diagrams to enhance parsimony.

**Mediated (and indirect effects) models.** The third set of analyses tested the role of perceived task values as mediators or indirect paths between initial interest and goals and the final outcome measures. First, we established that the predictors were significantly related to task values by using the direct effects regression model to predict intrinsic and utility value. Second, we established that the mediators were related to the outcomes, while controlling for the original predictors, by modifying the direct effects model to include task values. Therefore, we added intrinsic and utility value to the direct effects model, as well as all possible two- and three-way interactions between task values, initial interest, and goals. For each set of analyses, only significant interactions were retained in the model. Thus, unless otherwise noted, the mediation model contained 12 terms: 10 terms from the direct

effects model with the addition of the 2 value terms. Finally, we tested the significance of mediated and indirect effects as in Study 1 (Kenny et al., 1998).

**Direct effects on utility value.** The significant main effects of initial interest,  $t(193) = 3.39, p = .001$  ( $\beta = .25$ ), and mastery-approach goals,  $t(193) = 2.52, p = .01$  ( $\beta = .17$ ), indicated that students with higher levels of initial interest and mastery-approach goals were more likely to find utility value in the drills. No other effects were significant.

**Direct effects on intrinsic value.** The significant main effect of initial interest,  $t(192) = 5.01, p < .01$  ( $\beta = .35$ ), indicated that players with higher levels of initial interest were more likely to find intrinsic value in the drills. There was also a significant interaction between mastery- and performance-approach goals,  $t(192) = -3.46, p < .01$ , ( $\beta = -.23$ ). An inspection of the predicted values indicated that, for players with low levels of performance-approach goals, high levels of mastery-approach goals lead to more intrinsic value ( $\hat{Y} = 5.42$ ) than low levels of

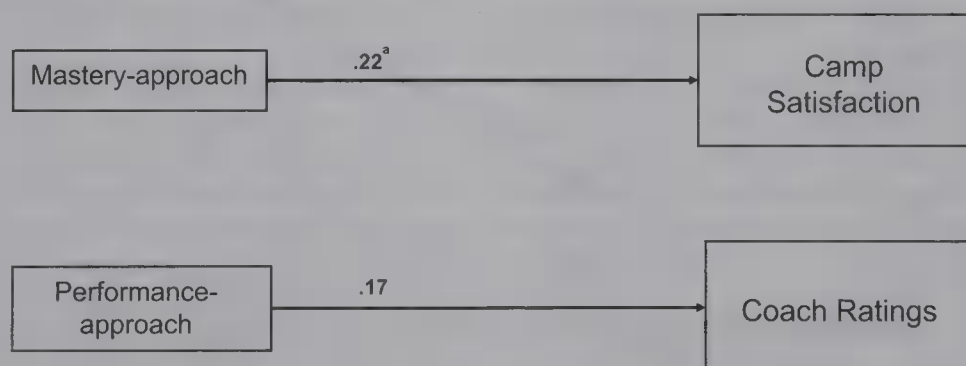


Figure 5. Study 2 path model of direct effects on football camp outcomes. All paths are standardized regression coefficients from simultaneous multiple regressions that included only the direct effects in the model. All paths are significant at  $p < .05$ . \*The direct path from mastery-approach goals to camp satisfaction is moderated by a three-way interaction between initial interest, mastery-approach goals, and performance-approach goals (see the text for statistics).



mastery-approach goals ( $\hat{Y} = 4.75$ ). In contrast, for players with high levels of performance-approach goals, high levels of mastery-approach goals did not lead to more intrinsic value ( $\hat{Y} = 5.01$ ) than for players with low levels of mastery-approach goals ( $\hat{Y} = 5.21$ ).

*Mediated and indirect effects on camp satisfaction.* In the mediation model, the direct effect for utility value was significant,  $t(154) = 3.04, p < .01$  ( $\beta = .22$ ), indicating that players who perceived the drills to be useful reported more satisfaction with camp. The direct effect for intrinsic value was significant,  $t(154) = 2.03, p = .04$  ( $\beta = .16$ ), indicating that players who perceived the drills to be enjoyable reported more satisfaction with camp. The direct effect for mastery-approach goals was significant,  $t(154) = 3.05, p < .01$  ( $\beta = .19$ ), but reduced in size from the direct effects model ( $\beta = .22$ ). The direct effect of initial interest was no longer significant,  $t(154) = 1.32, p = .19$  ( $\beta = .10$ ), and reduced in size from the direct effects model ( $\beta = .20$ ). The formal tests of mediation indicated that the relationship between initial interest and camp satisfaction was fully mediated by utility value ( $z = 2.60, p < .01$ ) and intrinsic value ( $z = 1.88, p = .06$ ). The relationship between mastery-approach goals and camp satisfaction was partially mediated by utility value ( $z = 1.94, p = .05$ ). In addition to these main effects, the three-way interaction with initial interest, mastery-approach, and performance-approach goals remained significant in the mediated model,  $t(154) = -5.38, p < .01$  ( $\beta = -.39$ ).

*Mediated and indirect effects on coach ratings.* The significant main effect of performance-approach goals,  $t(178) = 2.26, p = .03$ , ( $\beta = .17$ ) indicated that players who adopted performance-approach goals performed better and put more effort into camp than those who did not adopt performance-approach goals. In addition, the interaction between initial interest and utility value trended toward significance,  $t(178) = 1.73, p = .09$  ( $\beta = .15$ ). This interaction is graphed in Figure 6. The simple slopes analysis indicated that utility value predicted coach ratings when initial interest in football was high ( $\beta = .24, p = .07$ ), but not when

initial interest was low ( $\beta = -.04, p = .62$ ). The complete path model for camp satisfaction and coach ratings is presented in Figure 7.

### Structural Equation Model

In order to test the overall fit of the model, we used Lisrel 8.72 to simultaneously test the model presented in Figure 7. The results indicate that the overall model provided a satisfactory fit to the data,  $\chi^2(27, N = 155) = 747.96, p < .001$ , root-mean-square error of approximation = .08, comparative fit index = .90, standardized root-mean-square residual = .14.

### Discussion—Study 2

Overall, the results of Study 2 are consistent with the results of Study 1 and with prior research. Both intrinsic and utility task values uniquely predicted satisfaction with camp, providing further evidence that these task values contribute to individuals' motivation. Campers who perceived football drills as enjoyable and useful for developing skills finished camp with greater overall satisfaction than did those who found less value in these drills. Mastery-approach goals and initial interest also predicted subsequent satisfaction with camp, and these effects were partially mediated by intrinsic and utility values. Thus, perceptions of value operated as a key process through which subsequent interest (i.e., camp satisfaction) was developed, replicating the findings of Study 1.

In addition, utility value, but not intrinsic value, predicted campers' performance at camp among those with high football interest. In other words, those who entered camp with a high level of interest in football and who perceived football drills as being instrumental for skill development were rated by their coaches as putting forth more effort and performing better. These effects are

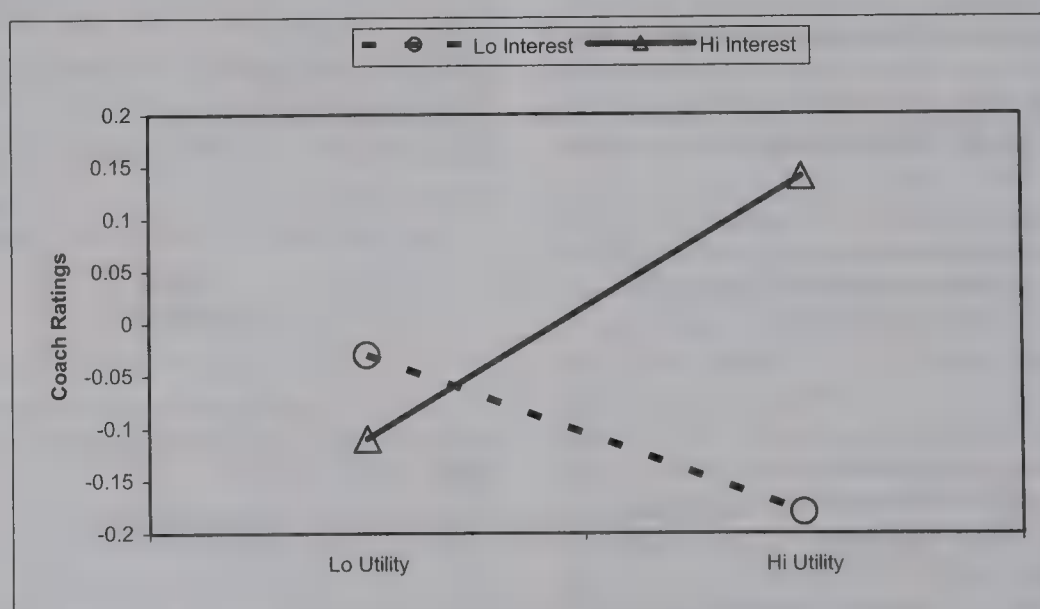


Figure 6. Interactive effects of initial interest and utility value on coach ratings of performance in Study 2. Interest = initial interest in football before camp started; Utility = usefulness of camp drills. The *Hi* and *Lo* prefixes indicate values at one standard deviation above and below the mean, respectively, for initial interest in football and utility value.

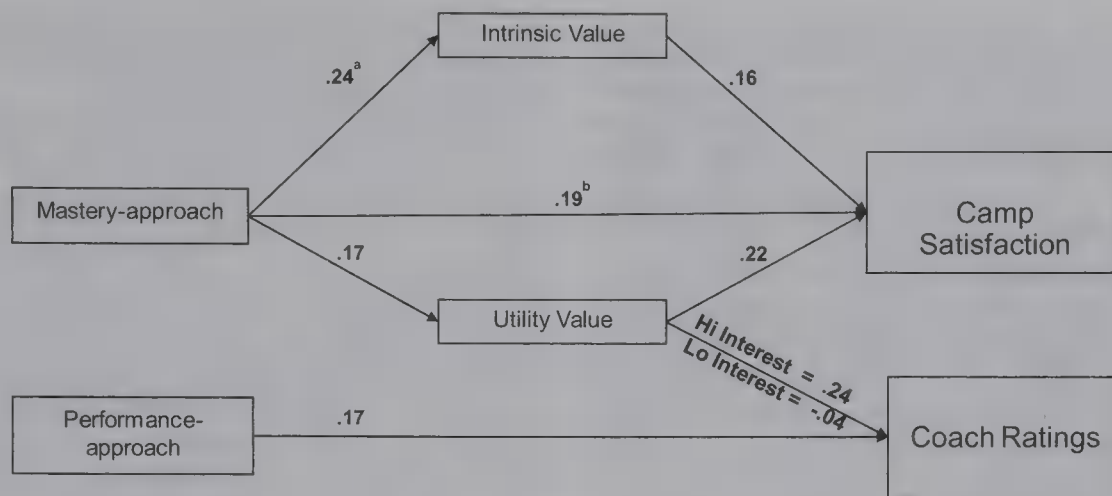


Figure 7. Study 2 path model of direct and mediated effects on football camp outcomes. All solid paths represent standardized regression coefficients from a simultaneous multiple regression that included all the variables in the model. All paths are significant at  $p < .05$  except the path from intrinsic value to camp interest ( $p = .06$ ), the path from initial interest to camp interest ( $p = .06$ ), and the interaction between initial interest and utility value on coach ratings ( $p = .085$ ). <sup>a</sup>The path from mastery-approach to intrinsic value is moderated by a significant interaction with performance-approach goals (see text for details). <sup>b</sup>The path from mastery-approach goals to camp satisfaction is moderated by a significant three-way interaction with initial interest and performance-approach goals (see the text for statistics).

comparable with the Study 1 finding that the perceived utility value of the course content predicted final grades.

The effects of achievement goals in this context were also comparable with the results of Study 1. The direct effects of mastery-approach goals on subsequent interest, and of performance-approach goals on performance, replicated the results of Study 1 as well as previous research (Harackiewicz, Barron, Pintrich, et al., 2002; Midgley et al., 2001), but within a sports context. This is important as there has been little work that assesses the effects of performance-approach goals on sports outcomes (e.g., Conroy et al., 2003; see Elliot, 2005, for a review). Moreover, as in Study 1, the effects of mastery-approach goals on subsequent interest were significant while we controlled for initial interest, suggesting that mastery goals play an important role in promoting subsequent interest in both contexts. Although initial interest and mastery goals were correlated,  $r(1150) = .61$  in Study 1, and  $r(237) = .20$  in Study 2, they each accounted for unique portions of the variance in subsequent interest. Thus, both studies provide strong support for the multiple goals perspective because both performance- and mastery-approach goals had positive effects on important outcomes.

One difference between the results obtained in the football camp and classroom studies was that mastery-approach goals did not predict intrinsic value in the sports context for everyone, whereas we did observe a main effect of mastery-approach goals on intrinsic value in the classroom. Rather, in the football camp, the effect of mastery-approach goals depended on the individual's level of performance-approach goals. Campers with high levels of either mastery- or performance-approach goals, but not both, perceived the most intrinsic value in camp drills. It is possible that pursuing one goal in the situation allowed the individual to find the enjoyable aspects of the drill, whereas pursuing both goals simultaneously prevented the individual from finding the enjoyable aspects of drills. In addition, the ratings of subsequent interest (i.e., camp satisfaction) were quite high in Study 2 ( $M = 6.33$  out of 7,

$SD = 0.75$ ) compared with Study 1 ( $M = 5.18$  out of 7,  $SD = 1.25$ ). Despite the potential for a ceiling effect in Study 2, achievement goals and task values were found to be significant predictors of subsequent interest, and these effects mirrored the pattern of findings found in Study 1.

Finally, Study 2 demonstrated the complexity of multiple goal effects (Barron & Harackiewicz, 2001, 2003). Not only did the effects of mastery goals depend on the level of performance goals, but these interactive goal effects also varied as a function of initial interest. For individuals with a lower initial interest in football, endorsing both goals was beneficial for subsequent interest. These individuals reported greater satisfaction with camp if they took advantage of the full range of achievement opportunities afforded them at camp, by focusing on both task mastery and performance relative to others. However, individuals who entered the camp situation with a higher initial interest in football benefited by adopting either a mastery-approach or performance-approach goal, but not both. One interpretation of this finding is that when individuals with elevated interest enter an achievement situation, it is motivationally adaptive for them to have one clear focus, or goal. These individuals are interested, paying attention, willing to engage in the task, and therefore need to focus on one particular objective. In contrast, when individuals with lower interest enter an achievement situation, it is motivationally adaptive to have multiple goals. These individuals are less interested and focused on the task, and therefore can benefit from many opportunities and objectives that can help them become task engaged. Put simply: Too many goals might frustrate and even distract the highly interested individual, whereas multiple goals might enable the less interested individual to remain engaged and motivated. It is unclear why this interaction appeared in Study 2 and not in Study 1. Future research will need to support this finding and its interpretations.

The relationship between performance-approach goals and subsequent interest found in Study 2 has not been found in classroom research (Harackiewicz, Barron, Pintrich, et al., 2002; Midgley et



al., 2001). The fact that we found such a relationship at a sports camp could be due to the enhanced salience of normative comparisons and competition. Performance-approach goals might predict subsequent interest in this context because competition and outperforming others is inherent to the game of football (Hidi & Harackiewicz, 2000; Kruglanski, 1975), and performance-approach goals are closely related to the object of the game. In the sports context, pursuing performance-approach goals may produce a more integrated type of self-regulation as comparisons with others are an inherent part of playing football (Deci & Ryan, 1985; Hidi & Harackiewicz, 2000). For example, a player has to be judged by the coaches to be better than other players at his position in order to play in the game. Consistent with this analysis, initial interest in football predicted both mastery-approach ( $\beta = .24$ ) and performance-approach goal adoption ( $\beta = .24$ ) to an equal degree. In contrast, the primary objective in college courses is to learn the material, and mastery-approach goals are most relevant to this objective. The results of Study 1 are consistent with this analysis as initial interest in the psychology class was a much stronger predictor of mastery-approach goals ( $\beta = .61$ ) than performance-approach goals ( $\beta = .11$ ). Considered together, the results of Studies 1 and 2 support the hypothesis that optimal motivation is a function of the match between an individual's adopted goals and the context (Harackiewicz, Barron, & Elliot, 1998; Harackiewicz & Sansone, 1991).

### General Discussion

The results of our research demonstrate the value of integrating multiple theoretical perspectives in the study of motivation. We demonstrated how achievement goals can operate as frameworks within which individuals perceive task value in activities, and how these task values may operate as proximal predictors of achievement outcomes. These conclusions are of course tempered by the correlational nature of our data; however, they do suggest potential pathways that can be further explored using causal methodology.

In addition, these results provide insight into the generalizability of motivational patterns across contexts. These classroom and sports camp contexts varied in terms of the type of activity (school vs. sports), participants' levels of prior experience with the activity (low vs. high), and age of participants (college vs. high school), respectively. In addition, we collected different measures of interest (course interest vs. camp satisfaction) and performance (course grades vs. coach ratings). Overall, there was remarkable consistency in the findings across the different measures in the two contexts. A comparison of the direct effects path models (Figures 2 and 5) and the mediated/indirect path models (Figures 3 and 7) reveals this similarity: Mastery-approach goals and initial interest both predicted subsequent interest, and these relationships were mediated by task values in both studies. Individuals in the classroom and on the playing field left the achievement situation with greater interest if they experienced intrinsic and utility value during task engagement. Moreover, performance-approach goals and utility value positively predicted performance in both studies, and the utility value effect revealed an indirect pathway from mastery-approach goals and initial interest to performance. Students and athletes both performed better when they found their respective tasks useful and personally meaningful. Adopting mastery-approach goals for the activity allowed students and ath-

letes to find intrinsic and utility value in the achievement tasks, and these values worked to promote both interest and performance in a classroom and on a football field.

Whereas little debate surrounds the association between intrinsic value and interest, some researchers have theorized that the relationship between utility value and interest is more complex (Eccles et al., 1983). Specifically, utility value has been described as sharing qualities with extrinsic motivation (Eccles & Wigfield, 2002), such that tasks that have high utility value are important not for the value inherent in doing them (Deci & Ryan, 1985), but for the value they have for engaging in other activities. Thus, the value is extrinsic to the immediate task, because it depends on its application to another situation (i.e., utility value is task extrinsic). Although many theorists have hypothesized that extrinsic motivators will impair the development of interest—whereas intrinsic, task-focused motivation will promote performance and interest (Ames, 1992; Deci & Ryan, 2000; Lepper & Henderlong, 2000)—the results of the current studies demonstrate that utility value is positively related to both motivation and performance. In addition, although task values and interest are correlated to a significant degree, the results of our confirmatory factor analyses demonstrate their empirical distinctiveness. This provides a foundation for continued research on the interrelationships of these unique—yet overlapping—constructs.

Some theorists have noted that utility value can be either intrinsic or extrinsic depending on why the task is perceived as useful (Malka & Covington, 2005; Simons, Vansteenkiste, Lens, & Lacante, 2004). For example, consider a football player who practices football drills during the off season in order to make the high school team. The immediate task (football drills) has utility value to the extent that it is instrumental in attaining a longer range goal (making the team). The utility of practicing football could be considered extrinsic if the player's parents (and not the player) want him to make the team (i.e., the goal is extrinsic). In contrast, the utility value could be considered intrinsic if the player has the personal goal of making the team (i.e., the goal is intrinsic). According to Simons, Vansteenkiste, Lens, and Lacante (2004), utility value can be intrinsic if its application is important to the person (i.e., person intrinsic) even if the value is not intrinsic to the task (i.e., task extrinsic). Research from this perspective has demonstrated that utility value, conceptualized as task extrinsic but person intrinsic, promotes motivation, interest, and performance compared with utility value that is person extrinsic (Simons, Dewitte, & Lens, 2000, 2003, 2004; Simons, Vansteenkiste, Lens, & Lacante, 2004; Vansteenkiste et al., 2004). Our measures of utility value emphasized the personal utility of the course content and football drills, and as such may be classified as person intrinsic. The positive relationship of utility value with important motivational outcomes in our studies provides support for the conceptualization of utility value as adaptive and beneficial (Brophy, 1999).

### *Implications for Multiple Goal Models, Limitations, and Future Research*

The generality of multiple goal effects on interest and performance was supported by our research. The pattern of achievement goal effects evidenced in previous classroom research was replicated in a sports context: Mastery-approach goals predicted inter-



est, and performance-approach goals predicted performance. In addition, these studies highlighted the intricacies of the multiple-goal perspective when considering optimal motivation (Barron & Harackiewicz, 2001, 2003; Harackiewicz, Barron, Pintrich, et al., 2002; Pintrich, 2003). In both the classroom and the football camp, mastery-approach goals were positively associated with subsequent interest when initial interest was high. However, when initial interest in football was low, mastery- and performance-approach goals worked synergistically to promote interest development. Clearly, more research is needed to further understand the relationship between multiple achievement goal adoption and interest.

In addition, now that there are more complete theoretical models of interest development (Hidi & Renninger, 2006), future research needs to test hypotheses about the contributions of task values and achievement goals in the development of interest. The current research provides insight into how these variables might work together, but further research is needed to continue to integrate these different motivational constructs and develop a more complete understanding of optimal motivation (Harackiewicz, Durik, & Barron, 2005; Pintrich, 2003).

The theoretical perspectives and constructs utilized in this research were extensions of our prior research integrating achievement goal and interest theory (Harackiewicz et al., 2000, 2008; Harackiewicz, Barron, Tauer, & Elliot, 2002; Harackiewicz et al., 2005; Hidi & Harackiewicz, 2000) and task value and interest theory (Durik, 2003; Durik & Harackiewicz, 2007; Godes, Hulleman, & Harackiewicz, 2007; Hulleman, Hendricks, & Harackiewicz, 2007). Our results are also consistent with other theoretical perspectives, such as the dual regulation systems proposed in the person-object-interaction model (e.g., Krapp, 2002, 2005), self-regulation research (Boekaerts, 2003), and models of volition (Kuhl, 2001). For example, in the person-object-interaction model (Krapp, 2005), the cognitive-emotional regulation system is responsible for developing connections between people and objects and activities, and this system has potential to add to researchers' understanding of situational interest. The cognitive subsystem, which is more reflective and conscious, could describe aspects of the perception of utility value, whereas the emotional subsystem, which is more reflexive and based on the person's immediate experience, could describe aspects of the perception of intrinsic value.

We acknowledge several limitations of this study. First, although we controlled for initial interest in the activities, we were not able to control for initial levels of performance in either of these studies. Controlling for initial levels of performance in our prior classroom research did not alter the effects of achievement goals on performance (Harackiewicz et al., 2000; Harackiewicz, Barron, Tauer, & Elliot; Harackiewicz et al., 2008), but without such controls in the present study, we do not know whether the effects of task value on performance would be impacted. Thus, future research needs to control for initial levels of performance, interest, and task values in order to make stronger claims regarding the causal influence of achievement goals and task values on educational outcomes (Maxwell & Cole, 2007). Second, our measure of subsequent interest in football camp reflected feelings of satisfaction as much as feelings of interest. Future research will need to replicate the interest findings in sports settings with measures that more closely parallel those used in the classroom.

Finally, although the research presented in this article highlights the important role that task values can play in the development of interest,

our results do not offer an explanation for how task values might facilitate the transition from situational interest to a more lasting form of personal interest. For example, if a student perceives utility value in psychology because it is instrumental in helping them gain admission into medical school, how does this task value lead to subsequent interest in psychology (rather than interest in medical school)? There are a number of possibilities that could be investigated. As interest researchers have suggested (e.g., Renninger, 2000), it is possible that finding meaning in psychology motivates students to spend more time studying and exploring psychology, which could then lead to a greater appreciation of the field and interest development. Additionally, students may become more engaged and involved with their learning, leading to feelings of task involvement and interest (Csikszentmihalyi, 1990; Dewey, 1913). These and other possible mechanisms responsible for the effects of task value on interest development warrant future exploration.

## References

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- Ames, C. (1984). Achievement attributions and self-instructions under competitive and individualistic goal structures. *Journal of Educational Psychology*, 76, 478–487.
- Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology*, 84, 261–271.
- Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology*, 80, 260–267.
- Austin, J. T., & Vancouver, J. B. (1996). Goal constructs in psychology: Structure, process, and content. *Psychological Bulletin*, 120, 338–375.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.
- Barron, K. E., & Harackiewicz, J. M. (2001). Achievement goals and optimal motivation: Testing multiple goal models. *Journal of Personality and Social Psychology*, 80, 706–722.
- Barron, K. E., & Harackiewicz, J. M. (2003). Revisiting the benefits of performance-approach goals in the college classroom: Exploring the role of goals in advanced college courses. *International Journal of Educational Research*, 38, 357–374.
- Boekaerts, M. (2003). Towards a model that integrates motivation, affect and learning. *British Journal of Educational Psychology, Monograph Series II*, 2, 173–189.
- Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology: An International Review*, 54, 199–231.
- Bong, M. (2001a). Between- and within-domain relations of academic motivation among middle and high school students: Self-efficacy, task-value, and achievement goals. *Journal of Educational Psychology*, 93, 23–34.
- Bong, M. (2001b). Role of self-efficacy and task-value in predicting college students' course performance and future enrollment intentions. *Contemporary Educational Psychology*, 26, 553–570.
- Brophy, J. (1999). Toward a model of the value aspects in education: Developing appreciation for particular learning domains and activities. *Educational Psychologist*, 34, 75–85.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.
- Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2006, April). *Predicting student achievement for low stakes tests with effort and task value*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- Conroy, D. E., Elliot, A. J., & Hofer, S. M. (2003). A 2 × 2 achievement



- goals questionnaire for sport: Evidence for factorial invariance, temporal stability, and external validity. *Journal of Sport & Exercise Psychology*, 25, 456–476.
- Csikszentmihalyi, M. (1990). *Finding flow: The psychology of engagement with everyday life*. New York: Basic Books.
- DeBacker, T. K., & Nelson, R. M. (1999). Variations on an expectancy-value model of motivation in science. *Contemporary Educational Psychology*, 24, 71–94.
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum Press.
- Deci, E. L., & Ryan, R. M. (2000). The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11, 227–268.
- Dewey, J. (1913). *Interest and effort in education*. Cambridge, MA: Riverside Press.
- Duda, J. (1995). Motivation in sports settings: A goal perspective approach. In G. C. Roberts (Ed.), *Motivation in sport and exercise* (pp. 57–91). Champaign, IL: Human Kinetics Books.
- Durik, A. M. (2003). Personal and situational factors involved in the development of interest (Doctoral dissertation, University of Wisconsin—Madison). *Dissertation Abstracts International*, 64(8-B), 2004.
- Durik, A. M., & Harackiewicz, J. M. (2007). Different strokes for different folks: How individual interest moderates the effects of situational factors on task interest. *Journal of Educational Psychology*, 99, 597–610.
- Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist*, 41, 1040–1048.
- Eccles, J., Adler, T. F., Futterman, R., Goff, S. B., Kaczala, C. M., Meece, J. L., & Midgley, C. (1983). Expectancies, values, and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives: Psychological and sociological approaches* (pp. 75–146). San Francisco: Freeman.
- Eccles, J. S. (2005). Subjective task value and the Eccles et al. model of achievement-related choices. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 105–121). New York: Guilford.
- Eccles, J. S., & Harold, R. D. (1991). Gender differences in sport involvement: Applying the Eccles’ expectancy-value model. *Journal of Applied Sport Psychology*, 3, 7–35.
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53, 109–132.
- Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist*, 34, 169–189.
- Elliot, A. J. (2005). A conceptual history of the achievement goal construct. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 52–72). New York: Guilford.
- Elliot, A. J., & McGregor, H. (2001). A 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology*, 80, 501–519.
- Elliot, A. J., McGregor, H., & Gable, S. (1999). Achievement goals, study strategies, and exam performance: A mediational analysis. *Journal of Experimental Social Psychology*, 91, 549–563.
- Flum, H., & Kaplan, A. (2006). Exploratory orientation as an educational goal. *Educational Psychologist*, 41, 99–110.
- Godes, O., Hulleman, C. S., & Harackiewicz, J. M. (2007, April). *Boosting students’ interest in math with utility value: Two experimental tests*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Grant, H., & Dweck, C. S. (2003). Clarifying achievement goals and their impact. *Journal of Personality and Social Psychology*, 85, 541–553.
- Harackiewicz, J. M., Barron, K. E., & Elliot, A. J. (1998). Rethinking achievement goals: When are they adaptive for college students and why? *Educational Psychologist*, 33, 1–21.
- Harackiewicz, J. M., Barron, K. E., Pintrich, P. R., Elliot, A. J., & Thrash, T. M. (2002). Revision of achievement goal theory: Necessary and illuminating. *Journal of Educational Psychology*, 94, 638–645.
- Harackiewicz, J. M., Barron, K. E., Tauer, J. M., Carter, S. M., & Elliot, A. J. (2000). Short-term and long-term consequences of achievement goals: Predicting interest and performance over time. *Journal of Educational Psychology*, 92, 316–330.
- Harackiewicz, J. M., Barron, K. E., Tauer, J. M., & Elliot, A. J. (2002). Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of Educational Psychology*, 94, 562–575.
- Harackiewicz, J. M., Durik, A. M., & Barron, K. E. (2005). Multiple goals, optimal motivation, and the development of interest. In S. M. Laham, J. P. Forgas, & K. D. Williams (Eds.), *Social motivation: Conscious and unconscious processes* (pp. 21–39). New York: Cambridge University Press.
- Harackiewicz, J. M., Durik, A. M., Barron, K. E., Linnenbrink-Garcia, E. A., & Tauer, J. M. (2008). The role of achievement goals in the development of interest: Reciprocal relations between achievement goals, interest and performance. *Journal of Educational Psychology*, 100, 105–122.
- Harackiewicz, J. M., & Sansone, C. (1991). Goals and intrinsic motivation: You can get there from here. In M. L. Maehr & P. R. Pintrich (Eds.), *Advances in motivation and achievement* (Vol. 7, pp. 21–49). Greenwich, CT: JAI Press.
- Heckhausen, J. (Ed.). (2000). *Motivational psychology of human development*. London: Elsevier.
- Hidi, S. (1990). Interest and its contribution as a mental resource for learning. *Review of Educational Research*, 60, 549–571.
- Hidi, S. (2000). An interest researcher’s perspective: The effects of extrinsic and intrinsic factors on motivation. In J. M. Harackiewicz & C. Sansone (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 309–339). San Diego, CA: Academic Press.
- Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research*, 70, 151–179.
- Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, 41, 111–127.
- Hulleman, C. S., Hendricks, B. L., & Harackiewicz, J. M. (2007, April). *The role of utility value in promoting classroom interest*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Jöreskog, K. G., & Sörbom, D. (1998). *LISREL 8: Structural equation modeling with the SIMPLIS command language* [Computer software manual]. Chicago: Scientific Software International.
- Kenny, D. A., Kashy, D. A., & Bolger, N. (1998). Data analysis in social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (4th ed., pp. 233–265). New York: McGraw-Hill.
- Krapp, A. (2002). An educational-psychological theory of interest and its relation to SDT. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 405–427). Rochester, NY: University of Rochester Press.
- Krapp, A. (2005). Basic needs and the development of interest and intrinsic motivational orientations. *Learning and Instruction*, 15, 381–395.
- Kruglanski, A. W. (1975). The endogenous–exogenous partition in attribution theory. *Psychological Review*, 82, 387–406.
- Kuhl, J. (2001). A functional approach to motivation. In A. Efklides, J. Kuhl, & R. M. Sorrentino (Eds.), *Trends and prospects in motivation research* (pp. 239–268). Dordrecht, The Netherlands: Kluwer.
- Lee, F. K., Sheldon, K. M., & Turban, D. B. (2003). Personality and the goal-striving process: The influence of achievement goal patterns, goal level, and mental focus on performance and enjoyment. *Journal of Applied Psychology*, 88, 256–265.
- Lepper, M. R., & Henderlong, J. (2000). Turning “play” into “work” and “work” into “play”: 25 years of research on intrinsic versus extrinsic motivation. In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 257–307). San Diego, CA: Academic Press.
- Linnenbrink, E. A. (2005). The dilemma of performance-approach goals:

- The use of multiple goal contexts to promote students' motivation and learning. *Journal of Educational Psychology*, 97, 197–213.
- Linnenbrink, E. A., & Pintrich, P. R. (2000). Multiple pathways to learning and achievement: The role of goal orientation in fostering adaptive motivation, affect, and cognition. In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 195–227). San Diego, CA: Academic Press.
- Mac Iver, D. J., Stipek, D. J., & Daniels, D. H. (1991). Explaining within-semester changes in student effort in junior high school and senior high school courses. *Journal of Educational Psychology*, 83, 201–211.
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, 7, 83–104.
- Maehr, M. (1989). Thoughts about motivation. In C. Ames & R. Ames (Eds.), *Research on motivation in education: Goals and cognitions* (Vol. 3, pp. 299–315). New York: Academic Press.
- Malka, A., & Covington, M. V. (2005). Perceiving school performance as instrumental to future goal attainment: Effects on graded performance. *Contemporary Educational Psychology*, 30, 60–80.
- Markus, H. (1977). Self-schemata and processing information about the self. *Journal of Personality and Social Psychology*, 35, 63–78.
- Maxwell, S. E., & Cole, D. A. (2007). Bias in cross-sectional analyses of longitudinal mediation. *Psychological Methods*, 12, 23–44.
- Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational Psychology*, 82, 60–70.
- Midgley, C., Kaplan, A., & Middleton, M. (2001). Performance-approach goals: Good for what, for whom, under what circumstances, and at what cost? *Journal of Educational Psychology*, 93, 77–86.
- Mitchell, M. (1993). Situational interest: Its multifaceted structure in the secondary school mathematics classroom. *Journal of Educational Psychology*, 85, 424–436.
- Muthén, L. K., & Muthén, B. O. (2006). *Mplus user's guide* (4th ed.) [Computer software manual]. Los Angeles: Authors.
- Nicholls, J. G. (1979). Quality and equality in intellectual development: The role of motivation in education. *American Psychologist*, 34, 1071–1084.
- Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological Review*, 91, 328–346.
- Pintrich, P. R. (2000a). An achievement goal theory perspective on issues in motivation terminology, theory, and research. *Contemporary Educational Psychology*, 25, 92–104.
- Pintrich, P. R. (2000b). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology*, 92, 544–555.
- Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology*, 95, 667–686.
- Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research, and applications* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.
- Rawsthorne, L. J., & Elliot, A. J. (1999). Achievement goals and intrinsic motivation: A meta-analytic review. *Personality and Social Psychology Review*, 3, 326–344.
- Renninger, K. A. (2000). Individual interest and its implications for understanding intrinsic motivation. In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 373–404). San Diego, CA: Academic Press.
- Renninger, K. A., & Hidi, S. (2002). Student interest and achievement: Developmental issues raised by a case study. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 173–195). San Diego, CA: Academic Press.
- Rubin, D. R. (1976). Inference and missing data. *Biometrika*, 63, 581–592.
- Schiefele, U. (1991). Interest, learning, and motivation. *Educational Psychologist*, 26, 299–323.
- Schiefele, U. (1999). Interest and learning from text. *Scientific Studies of Reading*, 3, 257–279.
- Senko, C., & Harackiewicz, J. M. (2005). Regulation of achievement goals: The role of competence feedback. *Journal of Educational Psychology*, 97, 320–336.
- Shah, J. Y., & Kruglanski, A. (2000). The structure and substance of intrinsic motivation. In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 105–127). San Diego, CA: Academic Press.
- Shim, S., & Ryan, A. (2005). Changes in self-efficacy, challenge avoidance, and intrinsic value in response to grades: The role of achievement goals. *Journal of Experimental Education*, 73, 333–349.
- Simons, J., Dewitte, S., & Lens, W. (2000). Wanting to have versus wanting to be: The effect of perceived instrumentality on goal orientation. *British Journal of Psychology*, 91, 335–351.
- Simons, J., Dewitte, S., & Lens, W. (2003). "Don't do it for me. Do it for yourself!" Stressing the personal relevance enhances motivation in physical education. *Journal of Sport and Exercise Psychology*, 25, 145–160.
- Simons, J., Dewitte, S., & Lens, W. (2004). The role of different types of instrumentality in motivation, study strategies, and performance: Know why you learn, so you'll know what you learn! *British Journal of Educational Psychology*, 74, 343–360.
- Simons, J., Vansteenkiste, M., Lens, W., & Lacante, M. (2004). Placing motivation and future time perspective theory in a temporal perspective. *Educational Psychology Review*, 16, 121–139.
- Updegraff, K. A., Eccles, J. S., Barber, B. L., & O'Brien, K. M. (1996). Course enrollment as self-regulatory behavior: Who takes optional high school math courses? *Learning and Individual Differences*, 8, 239–259.
- Vansteenkiste, M., Simons, J., Lens, W., Soenens, B., Matos, L., & Lacante, M. (2004). Less is sometimes more: Goal content matters. *Journal of Educational Psychology*, 96, 755–764.
- Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, 92, 548–573.
- Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review*, 6, 49–78.
- Wigfield, A., & Eccles, J. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review*, 12, 265–310.
- Wigfield, A., & Eccles, J. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In A. Wigfield & J. Eccles (Eds.), *Development of achievement motivation* (pp. 91–120). San Diego, CA: Academic Press.
- Wolters, C. A., Yu, S. L., & Pintrich, P. R. (1996). The relation between goal orientation and students' motivational beliefs and self-regulated learning. *Learning and Individual Differences*, 8, 211–238.
- Xiang, P., Chen, A., & Bruene, A. (2005). Interactive impact of intrinsic motivators and extrinsic rewards on behavior and motivation outcomes. *Journal of Teaching in Physical Education*, 24, 179–197.
- Xiang, P., McBride, R., & Bruene, A. (2004). Fourth graders' motivation in an elementary physical education running program. *Elementary School Journal*, 104, 253–266.
- Zimmerman, B. J., & Kitsantas, A. (1997). Developmental phases in self-regulation: Shifting from process goals to outcome goals. *Journal of Educational Psychology*, 89, 29–36.



## Appendix

## Interest Items for Study 2

Scale	Items
Initial interest	1. Football is very important to me. 2. Football is my favorite sport to play. 3. During the summer I mostly prepare for football season. 4. I came to this camp because I thought it would be fun. 5. Football season is my favorite part of the year. 6. Football practice is my favorite part of the day.
Camp interest	1. I had fun at this camp. 2. Overall I am satisfied with this camp. 3. I am disappointed with this camp. 4. Instead of attending this camp, I wish I had worked out on my own. 5. Camp was a good use of my time.

Received April 16, 2007

Revision received October 4, 2007

Accepted October 7, 2007 ■



## AMERICAN PSYCHOLOGICAL ASSOCIATION

### SUBSCRIPTION CLAIMS INFORMATION

Today's Date: \_\_\_\_\_

We provide this form to assist members, institutions, and nonmember individuals with any subscription problems. With the appropriate information we can begin a resolution. If you use the services of an agent, please do **NOT** duplicate claims through them and directly to us. **PLEASE PRINT CLEARLY AND IN INK IF POSSIBLE.**

PRINT FULL NAME OR KEY NAME OF INSTITUTION \_\_\_\_\_

MEMBER OR CUSTOMER NUMBER (MAY BE FOUND ON ANY PAST ISSUE LABEL) \_\_\_\_\_

ADDRESS \_\_\_\_\_

DATE YOUR ORDER WAS MAILED (OR PHONED) \_\_\_\_\_

CITY \_\_\_\_\_

STATE/COUNTRY \_\_\_\_\_

ZIP \_\_\_\_\_

\_\_\_\_ PREPAID \_\_\_\_ CHECK \_\_\_\_ CHARGE

CHECK/CARD CLEARED DATE: \_\_\_\_\_

YOUR NAME AND PHONE NUMBER \_\_\_\_\_

(If possible, send a copy, front and back, of your cancelled check to help us in our research of your claim.)

ISSUES: \_\_\_\_ MISSING \_\_\_\_ DAMAGED

TITLE \_\_\_\_\_

VOLUME OR YEAR \_\_\_\_\_

NUMBER OR MONTH \_\_\_\_\_

 \_\_\_\_\_  
 \_\_\_\_\_  
 \_\_\_\_\_

 \_\_\_\_\_  
 \_\_\_\_\_  
 \_\_\_\_\_

 \_\_\_\_\_  
 \_\_\_\_\_  
 \_\_\_\_\_

*Thank you. Once a claim is received and resolved, delivery of replacement issues routinely takes 4-6 weeks.*

(TO BE FILLED OUT BY APA STAFF)

DATE RECEIVED: \_\_\_\_\_

DATE OF ACTION: \_\_\_\_\_

ACTION TAKEN: \_\_\_\_\_

INV. NO. &amp; DATE: \_\_\_\_\_

STAFF NAME: \_\_\_\_\_

LABEL NO. &amp; DATE: \_\_\_\_\_

Send this form to APA Subscription Claims, 750 First Street, NE, Washington, DC 20002-4242

**PLEASE DO NOT REMOVE. A PHOTOCOPY MAY BE USED.**

# Striving for Social Dominance Over Peers: The Implications for Academic Adjustment During Early Adolescence

Sarah M. Kiefer  
University of South Florida

Allison M. Ryan  
University of Illinois, Urbana-Champaign

This study investigated the proposal that social dominance goals are an important, but overlooked, aspect of social goals for young adolescents' academic adjustment. Self-reports of social goals (dominance, intimacy, and popularity goals) early in the school year were used to predict subsequent engagement (self-reports and peer nominations of effort toward school work and disruptive behavior) and achievement (i.e., grades) when students were in 6th grade ( $N = 718$ ) and again after the transition to middle school when students were in 7th grade ( $N = 656$ ; 52% African American and 48% White; 52% female and 48% male). In line with hypotheses, social dominance goals were associated with maladaptive forms of engagement and low achievement in 6th and 7th grades. For intimacy goals, relations were more limited, but when found, these goals were associated with adaptive forms of engagement in 6th and 7th grades. Popularity goals were not generally associated with engagement or achievement. The exception was 6th-grade African American girls, for whom popularity goals were associated with maladaptive engagement (i.e., low effort, high disruptive behavior, and low peer nominations for trying hard and getting good grades).

**Keywords:** social goals, motivation, achievement, engagement, young adolescents

Students' goals are important to understanding their motivation, behavior, and achievement in school (Elliot & Dweck, 2005; Wigfield, Eccles, Schiefele, Rosser, & Davis-Kean, 2006). Goals are cognitive representations of things students want to accomplish and provide direction and energy for behavior (Ames, 1992; Elliot, 1999; Pintrich, 2000; Urdan, 1997). Academic goals, in particular achievement goals, have been the focus of much research and have provided insight into students' behaviors and performance in academic settings (Elliot & Dweck, 2005). However, students pursue a wide range of goals beyond academic achievement goals. Given the social nature of school, it is not surprising that students' social goals are increasingly being appreciated as important to understanding engagement and achievement in school (Anderman, 1999; Dowson & McInerney, 2001; Horst, Finney, & Barron, in press; Ryan, Kiefer, & Hopkins, 2004; Urdan & Maehr, 1995; Wentzel, 2005). The implications of social goals for engagement and achievement in school have received far less attention than academic goals and psychologists' understanding of the nature and consequences of social goals is far from complete. A social goal that has been largely overlooked is the goal for dominance among peers. No studies to date have examined the implications of social

dominance goals for academic adjustment. This is surprising given that strivings for dominance have held a prominent position in much research and theorizing on motivation and personality (McAdams, 1988; McClelland, 1985; Murray, 1938; Veroff, 1957; Winter, 1973). Further, recent research has highlighted that aggressive and manipulative behavior among students is prevalent and important to social adjustment in school (Cillessen & Rose, 2005; Hawley, Little, & Rodkin, 2007). Thus, the aim of the present study is to investigate the implications of social dominance goals for academic engagement (i.e., effort toward school work and disruptive classroom behavior) and achievement (i.e., grades).

## Different Approaches Used to Examine Links Between Social Goals and Academic Adjustment

A variety of approaches have been used to investigate students' social goals in relation to academic adjustment. First, researchers have asked students how often they try to do various things (e.g., have fun, follow rules) to describe their social goals with regard to what they are trying to accomplish with peers (often referred to as a *content approach* to goals; Ford, 1992; Wentzel, 2000). With this approach, Wentzel (e.g., 1992, 1996) established that social responsibility and prosocial goals positively relate to academic engagement and achievement. Building on Wentzel's work, Anderman (1999) documented that popularity goals negatively relate to academic engagement and achievement (see also Ryan, Hicks, & Midgley, 1997). Second, social goals have been conceptualized as distinct orientations toward social competence and have been linked to adjustment in the classroom (an *achievement goal approach* encompassing development and demonstration goals; Dweck & Leggett, 1988; Erdley, Cain, Loomis, Dumas-Hines, & Dweck, 1997; Ryan et al., 2004). A third approach has been to focus on a hybrid of social-academic goals (e.g., the goal to do

---

Sarah M. Kiefer, Department of Psychological and Social Foundations, College of Education, University of South Florida; Allison M. Ryan, Department of Educational Psychology, University of Illinois, Urbana-Champaign.

This work was supported by a grant from the Campus Research Board at the University of Illinois, Urbana-Champaign. The data presented, the statements made, and the views expressed are solely the responsibility of the authors.

Correspondence concerning this article should be addressed to Sarah M. Kiefer, University of South Florida, 4202 East Fowler Avenue, EDU 162, Tampa, FL 33620-5650. E-mail: kiefer@coedu.usf.edu



well academically to secure social approval or the goal to do well academically to fulfill a social obligation; Dowson & McInerney, 2001; Urdan & Maehr, 1995). These various approaches highlight different aspects of social goals and provide insight into connections between social goals and academic adjustment for students in schools. However, absent from all of these approaches is attention to social dominance goals.

### Social Dominance Goals as Representative of One of Three Basic Motives

In his seminal work, *Human Motivation*, McClelland (1985) identified three fundamental motivational systems that energize and direct behavior: power, achievement, and affiliation. The power motive concerns the desire to have impact, control, or influence over another person or group of people. The achievement motive concerns the desire to succeed with a consistent focus on doing things better. The affiliative motive encompasses the general desire to be with others as well as the desire for intimate relationships characterized by warmth and disclosure of personal thoughts and feelings. Other theorists have also drawn attention to the need for power as an important aspect of human motivation and have distinguished it from the need for affiliation in social relationships (McAdams, 1988; Murray, 1938; Winter, 1973; Winter & Stewart, 1978).

McClelland (1985) conceptualized the power, achievement, and affiliation motives as enduring traits and distinguished these broader motive systems from more specific goal strivings. We conceptualize goals as the concrete and situation-specific representations of the more abstract and general motivational dispositions (Elliot, 1999). As the more situation-specific manifestations of motive systems, goals should be related to situation-specific behaviors of students (i.e., engagement and achievement in school). Thus, we focus on the social dominance goals that students pursue in their peer relationships at school. We do not examine the antecedents of students' social dominance goals but assume they stem in part from individual differences in students' power motive as well as features of the social context (e.g., school culture, teacher behaviors, peer dynamics, parental values, and neighborhood setting).

We define social dominance goals as a focus on having power over peers, characterized by getting peers to comply with their wishes and instilling fear in peers. We acknowledge that this is not the only type of social goal that may emanate from the power motive. For example, a more positive manifestation could be a social goal focused on leadership. We investigate social dominance goals because recent research has suggested they are prevalent and important in the lives of young adolescents. Manipulative, indirect, and direct aggressive behavior has been found to play a central role in peer dynamics and social success (Cillessen & Rose, 2005). For decades, any type of aggressive behavior was considered a form of social incompetence and a correlate and precursor of peer rejection (operationalized as many *don't like* nominations and few *like* nominations from peers). However, recent research has highlighted that, when social status is measured with indicators such as *popular* or *cool* nominations instead of *like* nominations, manipulative, indirect, and even direct aggressive behaviors are often associated with high social status (Cillessen & Mayeux, 2004; Cillessen & Rose, 2005; Rodkin, Farmer, Pearl, &

Van Acker, 2000; Rose, Swenson, & Waller, 2004). It is important to note that social dominance goals are not synonymous with aggressive behavior. Aggressive behavior is one possible consequence of social dominance goals, or striving for power over peers. In our investigation, we examine the consequences of social dominance goals for engagement (effort and disruptive behavior) and achievement in school, and we hypothesize that there are serious costs of social dominance goals for these indicators of academic adjustment.

To fully understand the implications of social dominance goals, we examine them in tandem with two other social goals that represent McClelland's (1985) major motive systems: intimacy and popularity goals. Intimacy goals refer to a focus on establishing intimate peer relationships characterized by mutual support and disclosure of thoughts and feelings (Anderman, 1999; Jarvinen & Nicholls, 1996). Popularity goals refer to a focus on establishing high social status characterized by visibility and prestige within the larger peer group at school (Anderman, 1999; Jarvinen & Nicholls, 1996; Ryan et al., 1997). We view these two social goals as stemming, in part, from individual differences in students' affiliative and achievement motives. Here, too, we acknowledge that these are not the only two social goals that may emanate from these larger motives. Particularly for the achievement motive within the social domain, success and doing better in peer relations at school could take different forms. Students may try to improve their social skills and relationships relative to their own past experiences (i.e., a social development goal), or students may try to be better compared with others (i.e., a social demonstration-approach goal; see Dweck & Leggett, 1988; Ryan & Shim, 2006a, 2006b). Popularity goals are conceptually similar but more narrow in scope than social demonstration-approach goals. We chose intimacy and popularity goals because (a) they represent concerns that have been shown to be important in adolescent peer relationships (e.g., Bukowski & Hoza, 1989; Cillessen & Mayeux, 2004; Eder, 1985; Grotzinger & Crick, 1996; Hawley, 2003; Merten, 1997; Savin-Williams & Berndt, 1990) and (b) they have been associated with academic engagement (e.g., Anderman, 1999; Ryan et al., 1997). Given that we are proposing that social dominance goals are an important but overlooked aspect of young adolescents' social goals in school, it is important to demonstrate that the effects of social dominance goals exist above and beyond other social goals that have already been shown to predict engagement and achievement in this age group.

### Hypotheses for Social Dominance, Intimacy, and Popularity Goals and Academic Adjustment

Much theory and research indicate that social and academic adjustment are intimately related at school (e.g., Ladd, Kochenderfer, & Coleman, 1997; Patrick, 1997; Ryan, 2001; Wentzel, 2005). Social goals have implications for academic adjustment through the social behaviors as well as the types of social relationships they promote (Wentzel, 2005). Social dominance goals are likely to be associated with maladaptive forms of engagement and low achievement. A focus on having power over and instilling fear in peers is likely to lead to disruptive and off-task behavior and low achievement. In contrast, intimacy goals are likely to be associated with adaptive forms of engagement and achievement. A focus on building mutually supportive and close friendships is likely to lead



to positive behaviors, such as helping, cooperating, and sharing with peers. In line with this view, prosocial goals have been found to be positively associated with effort in the classroom (Wentzel, 1993), and intimacy goals have been found to be positively associated with positive attitudes toward school (Anderman, 1999).

Popularity goals are not expected to promote positive engagement and high achievement. Although much extant research has documented a positive association between social and academic success at school (see Rubin, Bukowski, & Parker, 2006, for a review), this association is generally found with measures that assess likeability among peers. When social status is measured as popularity, the association is much weaker (LaFontana & Cillessen, 2002). Students do not view effort and good grades in school as promoting popularity (Gorman, Kim, & Schimmelbusch, 2002; Juvonen & Murdock, 1993, 1995). The pursuit of popularity goals has been found to be related to negative attitudes about school (Anderman, 1999) and disengagement in the form of help avoidance (Ryan et al., 1997). Thus, we expect popularity goals to be associated with disengagement and low achievement. However, we expect that the associations for social dominance goals are stronger, because, given the content of social dominance goals as focused on being powerful, tough, and feared, the realization of such goals is likely to have a stronger association with disruptive and off-task behavior.

### Gender and Ethnicity

We examine gender and ethnic differences in social goals. In a recent review of research on gender differences in peer relationships, Rose and Rudolph (2006) concluded that boys tend to endorse more agentic and status-oriented goals, whereas girls tend to endorse more connection-oriented goals. These findings regarding social goals are in line with research showing that boys are socialized to be more competitive and dominant and girls are socialized to be more nurturing and caring (Eccles, 1993; Ruble, Martin, & Berenbaum, 2006). Thus, we expect boys to endorse social dominance and popularity goals more than girls do, and we expect girls to endorse intimacy goals more than boys do. Rose and Rudolph did not note any ethnic or racial differences in social goals, likely because most of the research has been on predominantly White samples. A contribution of the present research is that we address this gap. We hypothesize that ethnicity is associated with young adolescents' social goals, particularly social dominance goals.

McClelland (1985) theorized that a loss of or threat to one's power in society leads to a compensatory increase in the power motive. Thus, members of groups that are marginalized or discriminated against desire and focus more on exerting power than those who have not had such adverse social experiences. In support of this viewpoint, Veroff, Depner, Kulka, and Douvan (1980) found that African American men had higher power strivings than White men. Theory on the identity development of adolescent African American males in urban contexts describes a similar process. In response to discrimination and feelings of invisibility, African American males assume a *cool pose* (Cunningham & Meunier, 2004; Majors, 1991; Majors & Billson, 1992; Stevenson, 2004). The cool pose is a set of language and behaviors intended to display toughness, strength, and detachment. There is a bravado or macho attitude that encompasses the refusal to express emotions

that would render one vulnerable (Majors & Billson, 1992). In line with these ideas, we hypothesize that African American students (particularly males) are higher in social dominance goals and lower in intimacy goals compared with White students. We explore ethnic differences in popularity goals as well but generate no hypothesis, because these theories are more closely linked to social dominance and intimacy goals. In addition to group differences in mean levels, we also examine group differences in the nature of the relations of social goals to academic adjustment. This provides insight into whether the consequences of social goals for academic adjustment are the same for all groups and thus provides a more complete understanding (see Rowe, Vazsonyi, & Flannery, 1994, for a similar approach to understanding group differences in self-esteem).

### School Context

We conducted our study in the context of the transition to middle school in urban, predominantly low-income schools. The transition to middle school is a major event in the lives of early adolescents, which involves students adjusting to a larger and different school culture with many new students and different rules and norms of social and academic behaviors (Eccles, 2004). Transition stress may be exacerbated in low-income schools that have fewer resources to create emotionally supportive and high-quality instructional climates (Leventhal & Brooks-Gunn, 2000). Social relationships change dramatically because new friendships need to be formed and one's place in the social hierarchy must be redefined. Students go from being the oldest and the biggest students in elementary school to being the youngest and smallest in a new middle school. The salience and consequences of social dominance goals may change as students move from elementary to middle school. Pellegrini and Bartini (2001) found that being dominant was associated with aggression in new but not well-established social networks. Thus, the negative outcomes of social dominance goals, particularly for disruptive behavior, may be even stronger in the new setting of middle school.

### Overview of the Present Research

In summary, the present research examined the proposal that social dominance goals would have negative implications for students' academic adjustment at school. We conceptualize adjustment as both engagement (the behavioral involvement and participation necessary for school success) and achievement. To fully understand the role of social dominance goals for engagement and achievement, we consider social dominance goals in tandem with intimacy and popularity goals to identify whether there are unique effects. We expect that intimacy goals are positively associated with engagement and achievement. We expect that popularity goals are negatively associated with engagement and achievement (although less so than social dominance goals). Consistent with our conceptualization that goals precede academic adjustment, we used a prospective longitudinal design to examine whether social goals (dominance, intimacy, and popularity) measured in the fall of the school year predicted engagement (effort and disruptive behavior) and achievement (grades) measured in the spring (a 6-month time span). We used a variety of measures to assess engagement and achievement: self-reported effort, self-reported



disruptive behavior, peer nominations for working hard and getting good grades, peer nominations for not following school rules, and teacher-assigned grades. With a sample that was approximately half African American and half White, we examined gender, ethnic, and Gender  $\times$  Ethnic differences in the mean level of social goals as well as in the relations of social goals to engagement and achievement. We collected data from students when they were in their last year of elementary school (sixth grade) and again when they were in their first year of middle school (seventh grade) to examine whether the hypothesized relations between social dominance goals and engagement and achievement were stronger in the middle school context.

## Method

### *Schools*

The data were collected as part of the University of Illinois Adolescent Transitions Project, which is a 2-year longitudinal study examining changes in academic and social adjustment across the transition to middle school. Participants attended 1 of 15 elementary schools when they were in sixth grade and moved into 1 of 3 middle schools when they were in seventh grade. These elementary schools served nonmetropolitan, small urban communities. The average rate of eligibility for free or reduced-fee lunch across elementary schools was 66%, with 14 of the 15 elementary schools having greater than 50% of their students eligible and 1 school reporting 38% of their students as eligible. The average mobility rate (average number of students who transferred in or out of the school during the academic year) across elementary schools was 22%. The majority (12) had mobility rates between 20% and 29%, whereas 1 had a rate of 35%, and 2 had rates between 9% and 10%. The elementary schools had an average of 50% African American students, 46% White students, and 4% from other ethnic groups (13 of the 15 elementary schools had between 40% and 60% of both African American and White students, whereas 2 schools had slightly different configurations: 72%, 20%, and 8% and 57%, 30%, and 13% African American, White, and other ethnic groups, respectively).

At the middle school level, the average rate of eligibility for free or reduced-fee lunch across middle schools was 59% (45%, 55%, and 77% of students were eligible in each of the three middle schools). The average mobility rate across the middle schools was 24% (24%, 23%, and 23% for each of the three middle schools). The middle schools had an average of 48% African American students, 48% White students, and 4% from other ethnic groups (configurations were: 46%, 53%, and 1%; 37%, 61%, and 2%; and 59%, 31%, and 10% African American, White, and other ethnic groups, respectively, at each of the three middle schools).

### *Procedure*

Letters describing the project were given to all students to take home to their parents 2 weeks prior to each data collection. Given the substantial mobility rates in these schools, we maximized the sample size by recruiting new students at each wave. If parents did not want their children to participate in the study, they were instructed to have their child return an attached form to the teacher, call the school, or call the researchers at the university number

provided on the letter. All teachers were given two copies of the letter for each student, and teachers checked with students that the letters were delivered at home. Less than 5% of the parents declined to have their child participate at any wave.

Surveys were administered to students in their classrooms. Instructions and items were read aloud while students read along and responded. Students were told that the purpose of the survey was to find out about students' beliefs and behaviors. They were also told that the survey was not a test and that there were not right or wrong answers. Students were assured that the information in the survey would be kept confidential. In addition, they were told that filling out the survey was voluntary and that if at any point they wanted to stop, they could do so. We visited the schools 1 additional day to administer make-up surveys for students who were absent for survey administration.

### *Participants*

Given that the analyses examined the extent to which social goals in the fall predicted engagement and achievement in the spring, students who did not take the survey at both times within each year were dropped (20%). Further, students who were not African American or White were dropped because there were so few that it was not possible to analyze ethnic differences for these students. These restrictions yielded a sample of 718 students for Year 1 (52% female, 48% male; 52% African American, 48% White) and a sample of 656 students for Year 2 (53% female, 47% male; 52% African American, 48% White).

### *Attrition*

There were 373 students who were in both the first and second years of this study (i.e., completed surveys at all waves). Attrition was due to two factors: school participation and mobility rates. Regarding school participation, we lost some students who went to a nonparticipating middle school and gained some students from nonparticipating elementary schools (in total, three elementary schools and one middle school chose not to participate). After these students were accounted for, the sample instability was comparable with the mobility rates reported by the state for these schools.

We analyzed missing data to see if there were any significant differences between Year 1 students who remained in the study and those who were lost. There were no significant differences in social goals, self-reported effort, or self-reported disruptive behavior. However, students retained in the study had higher grade point averages, more peer nominations as students who try hard and get good grades, and fewer peer nominations as students who do not follow the rules compared with the students who were lost ( $t_s > 2.5$ ,  $p_s < .01$ ). Thus, in general, students who were retained had a more adaptive profile regarding academic engagement and achievement, which makes sense given that mobility is often due to financial strain and job instability within the family (Heinlein & Shinn, 2000). We also compared returning students with new students in Year 2 of the study but found no significant differences between these two groups.

### *Measures*

*Social goals.* Dominance, intimacy, and popularity goals were measured with items from the Social Goals Questionnaire (Jarvi-

nen & Nicholls, 1996). Given that the Social Goals Questionnaire was developed for and administered to high school students, we reviewed the items and chose the four items for each scale that were the most appropriate for sixth- and seventh-grade students. Social dominance goal items concern having power over peers. Intimacy goal items concern establishing intimate peer relationships. Popularity goal items concern establishing high social status. The items for each of the three social goal scales are displayed in Table 1. All items were rated on a 5-point scale (1 = *not at all true of me*, 3 = *somewhat true*, and 5 = *very true of me*). All items include the stem "I like it when," which focuses students on outcomes that would make them happy or feel successful. This is a unique feature to Nicholls's goal measures (e.g., Nicholls, Cheung, Lauer, & Pataschnick, 1989) compared with other measures of goals that use a range of stems such as "My goal is," "I try to," or "I want to" in addition to "I like it when" (e.g., Anderman, 1999). A study in the academic domain provided the validity information for this approach by documenting strong correlations ( $r > .63$ ) between analogous goal constructs with Nicholls's measure and another widely used measure, the Patterns of Adaptive Learning Survey (Midgley et al., 1998).

Because the items had not been administered to this age sample before, we conducted an exploratory factor analysis to ensure the social dominance, intimacy, and popularity goals were distinct constructs. Factor analysis is presented in the Results section and displayed in Table 1. Each of the three social goals was reliable in our sample in Year 1 (.83, .75, and .73 for dominance, popularity, and intimacy goals, respectively) and in Year 2 (.85, .76, and .79 for dominance, popularity, and intimacy goals, respectively). Social goals were fairly stable across the transition to middle school (fall of sixth grade to fall of seventh grade;  $r_s = .43, .46$ , and  $.45$  for dominance, intimacy, and popularity goals, respectively).

*Self-reports of classroom engagement.* Students reported on two facets of engagement in the classroom: effort and disruptive behavior. Items from the Rochester Assessment of Intellectual and Social Engagement were used to assess effort in school (see Miserandino, 1996; Skinner & Belmont, 1993). Sample items are

"I try very hard in school" and "I listen carefully in class." Students' disruptive behavior in class was assessed with a measure developed by Kaplan (e.g., Kaplan & Maehr, 1999). Sample items are "I disturb the lesson in class" and "I get into trouble in class." All items were rated on a 5-point scale (1 = *not at all true of me*, 3 = *somewhat true*, and 5 = *very true of me*). Each of the scales had four items, and all were found to be reliable in our sample at Year 1 ( $\alpha_s = .77$  and  $.74$  for effort and disruptive behavior, respectively) and at Year 2 ( $\alpha_s = .80$  and  $.80$  for effort and disruptive behavior, respectively). Engagement was fairly stable across the transition to middle school (spring of sixth grade to spring of seventh grade;  $r_s = .42$  and  $.53$  for effort and disruptive behavior, respectively).

The validity of the Rochester Assessment of Intellectual and Social Engagement has been demonstrated in research showing concordance between student and teacher reports of student effort (Skinner & Belmont, 1993). The validity of the disruptive behavior measure has been demonstrated in research showing that the more children reported their behavior as disruptive, the more official discipline referrals the children received (Kaplan & Maehr, 1999). In addition, children's reports of disruptive behavior, as assessed with this measure, are lower when teachers use effective management practices than when teachers do not do so (Patrick, Turner, Meyer, & Midgley, 2003).

*Peer nominations of classroom engagement.* We used Graham, Taylor, and Hudley's (1998) peer nomination measures in which students were asked, "Which students in your class work hard and get good grades?" and "Which students in your class do not follow school rules?" Students could nominate up to five same- or cross-gender peers in their grade for each question. Scores for peer nominations were calculated from the quotient of the number of nominations received by a student for a question over the number of classmates participating. A log transformation (with a small constant added to remove zero values) was applied to the peer scores to address the fact that peer nominations were not normally distributed, and then the peer scores were standardized by gender (see Rodkin et al., 2000). Peer nomination scores were fairly stable across the transition to middle

Table 1  
*Exploratory Factor Analyses of Social Goals Items*

Item	Dominance	Popularity	Intimacy
When I'm with people my own age I like it when . . .			
They worry that I'll hurt them	.85		
They are afraid of me	.84		
They know I'm tougher than them	.76		
I make them do what I want	.62		
I'm the most popular		.81	
I'm the coolest		.80	
They like me better than anyone else		.69	
Everyone wants me for a friend		.57	.37
Someone understands how I feel			.78
I really know someone's feelings			.73
I can make them happy			.72
I go out of my way to help them			.70
Eigenvalue	4.01	2.50	0.92
Percentage of variance explained	33.43	20.79	7.69
Cronbach's $\alpha$	0.83	0.75	0.73

Note. Loadings  $\geq .35$  are displayed. Principal axis factor analysis with oblimin rotation is shown.



school (spring of sixth grade to spring of seventh grade;  $r_s = .32$  and  $.45$  for works hard and gets good grades and does not follow school rules, respectively).

**Achievement.** Students' spring semester grades in reading, math, science, English, and social studies were collected from their school records. The grades were coded F = 1 through A+ = 13. The overall spring semester grade point average was computed by taking the mean of the five subject grades. Grade point average was fairly stable across the transition to middle school (spring semester of sixth grade to spring semester of seventh grade,  $r = .67$ ).

Results

Exploratory Factor Analyses

A principal axis factor analysis with oblimin rotation was conducted on the 12 social goal items at Wave 1. Factors whose eigenvalues were greater than 1 were extracted. The analysis yielded three factors, which accounted for 61.91% of the total variance (see Table 1). The three factors corresponded to the three social goals: dominance, intimacy, and popularity goals. All factor loadings were above .55 on their primary factor. No items loaded on another factor at greater than .40. The factor analysis was run again with varimax rotation and similar results were found.

Preliminary Analyses of Means and Correlations

The means and standard deviations are shown in Table 2. The range for all three social goals was between 1 and 5 for both Years 1 and 2. Median values for social goals were as follows: dominance goals (Year 1 = 1.50, Year 2 = 1.50), popularity goals (Year 1 = 2.86, Year 2 = 2.71), and intimacy goals (Year 1 = 4.00, Year 2 = 3.75). To determine if there were statistically significant gender and ethnic differences in social goals, two-way analysis of variance procedures were conducted. As expected, there were gender differences in social goals. For both Years 1 and 2, boys endorsed social dominance goals more than did girls: for Year 1,  $F(1, 763) = 49.45, p < .001$ ; for Year 2,  $F(1, 758) = 90.52, p < .001$ . In addition, for both Years 1 and 2, boys endorsed popularity goals more than did girls: for Year 1,  $F(1, 759) = 15.83, p < .001$ ; for Year 2,  $F(1, 756) = 21.96, p < .001$ . Girls

endorsed intimacy goals more than did boys: for Year 1,  $F(1, 760) = 69.47, p < .001$ ; for Year 2,  $F(1, 757) = 69.71, p < .001$ . There were no Gender  $\times$  Ethnicity interactions, indicating that these gender differences were the same for both African American and White students. As expected, there were also ethnic differences in social goals. For Year 1, African American students endorsed social dominance goals and popularity goals more than did White students,  $F(1, 763) = 26.42, p < .001$ ;  $F(1, 759) = 4.63, p < .05$ , respectively. For Year 2, African American students endorsed social dominance goals more than did White students,  $F(1, 758) = 18.80, p < .001$ , whereas White students endorsed intimacy goals more than did African American students,  $F(1, 757) = 5.84, p < .05$ .

The correlations among variables for the entire sample are shown in Table 2. An expected pattern of correlations was found among the variables. For the most part, social dominance goals were associated with subsequent disengagement and underachievement in sixth grade and seventh grade. Popularity goals were also associated with subsequent disengagement and underachievement, but the pattern was weaker. In contrast, intimacy goals were positively associated with most measures of subsequent engagement and achievement. We do not present correlations broken down by gender and ethnicity, because we are most interested in unique effects of goals (i.e., controlling for other social goals, gender, and ethnicity). Thus, we examined whether the relation of social goals with engagement and achievement varied by gender and ethnicity in the multivariate regression models.

Multiple Regression Analyses Examining Social Goals, Engagement, and Achievement in Sixth Grade (Elementary School)

We conducted separate regression analyses for the five engagement and achievement variables in sixth grade (students' last year of elementary school; see Table 3). The social goals (measured in the fall) were entered with gender and ethnicity. Gender and ethnicity were dichotomous variables with 1 coded as female and African American, respectively. Preliminary analyses tested interactions among social goals, gender and ethnicity. For example, we tested Dominance Goal  $\times$  Gender, Dominance Goal  $\times$  Ethnicity,

Table 2  
Correlations Among Social Goals and Academic Adjustment for the Overall Population

Variable	1	2	3	4	5	6	7	8
1. Dominance goals	—	.63**	-.18***	-.18**	.26**	-.21**	-.13**	.24**
2. Popularity goals	.64**	—	.08*	-.09*	.18**	-.05	.03	.10**
3. Intimacy goals	-.24**	.01	—	.21**	-.11**	.18**	.15**	-.12**
4. Effort	-.18**	-.10**	.26**	—	-.55**	.20**	.21**	-.21**
5. Disruptive behavior	.20**	.09*	-.10**	-.55**	—	-.31**	-.27**	.29**
6. Grade point average	-.27**	-.09*	.12**	.30**	-.32**	—	.57**	-.29**
7. PN: Effort/grades	-.17**	-.04	.11**	.27**	-.32**	.56**	—	-.20**
8. PN: Doesn't follow rules	.20**	.08*	-.13**	-.33**	.42**	-.37**	-.24**	—
M (SD) Year 1	2.05 (1.18)	2.97 (1.10)	3.86 (0.97)	3.80 (0.91)	2.82 (1.09)	7.98 (2.67)	3.01 (4.72)	2.42 (3.96)
M (SD) Year 2	1.80 (1.00)	2.77 (1.04)	3.64 (0.98)	3.77 (0.81)	2.74 (1.04)	6.44 (3.10)	2.31 (3.74)	1.34 (3.00)

Note. Year 1 (sixth grade) is below the diagonal (social goals in the fall and adjustment in the spring), and Year 2 (seventh grade) is above the diagonal (social goals in the fall and adjustment in the spring). PN = peer nomination.  
\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

Table 3  
Standardized Regression Coefficients for Predicting Academic Adjustment in the Spring From Social Goals in the Fall (Sixth Grade)

Variable in equation	Adjustment variable				
	Effort ( $\beta$ )	Disruptive ( $\beta$ )	GPA ( $\beta$ )	PN: Effort/grades ( $\beta$ )	PN: Not following rules ( $\beta$ )
Dominance goals	-.08	.17**	-.26***	-.09	.25**
Popularity goals	-.07	-.04	.07	-.21 <sup>†</sup>	-.15*
Intimacy goals	.20**	.01	.08	.01	.03
Gender	.19**	-.25***	-.05	.01	-.16 <sup>†</sup>
Ethnicity	.07	.14*	-.43***	-.32***	.30***
Gender $\times$ Ethnicity	-.10	.09	.16*	.16 <sup>†</sup>	-.25**
Dominance $\times$ Gender					.14*
Popularity $\times$ Gender	.11	.00		.34**	
Dominance $\times$ Ethnicity					-.19*
Popularity $\times$ Ethnicity	.07	-.15*		.23*	
Popularity $\times$ Gender $\times$ Ethnicity	-.19*	.18*		-.38***	
$R^2$	.11***	.11***	.21***	.11***	.16***

Note. For gender coding, 1 = female; for ethnicity coding, 1 = African American. GPA = grade point average; PN = peer nomination.

<sup>†</sup>  $p < .06$ . \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

and Dominance Goal  $\times$  Gender  $\times$  Ethnicity. We used procedures outlined by Aiken and West (1991) to test and interpret interactions. Main effect terms were standardized before computing interaction terms to avoid multicollinearity and aid interpretation of beta coefficients. Significance of interactions was determined if the  $R^2$  increased by a significant amount and if the beta coefficient for the interaction term was significant as well. Only significant interaction terms were retained in the final models (thus, the final models varied for different dependent variables). To interpret the significant interactions, we calculated the predicted values with unstandardized regression coefficients and conducted simple slope tests. Graphs were created with the predicted values of outcome variables for the students with goal scores one standard deviation above and below the mean.

**Social dominance goals.** As expected, social dominance goals were positively associated with subsequent disruptive behavior as well as peer nominations for not following the school rules. Further, social dominance goals were negatively associated with subsequent achievement. Thus, striving to be dominant and powerful among peers in the fall of the school year foreshadowed disruptive and disobedient behavior later in the year, according to both self-reports and peer nominations. Consistent with this pattern, a focus on dominance in peer relationships early in the school year was also associated with lower levels of achievement later in the school year. Social dominance goals were not associated with subsequent effort or peer nominations for effort/gets good grades.

There was a significant Social Dominance Goals  $\times$  Gender interaction for peer nominations of not following the rules, which indicated the positive association was stronger for girls compared with boys (see Figure 1). There was also a significant Social Dominance Goals  $\times$  Ethnicity interaction for peer nominations of not following the rules, which indicated a different relationship for White and African American students. Specifically, there was a positive association for social dominance goals and peer nominations of not following the rules for White students, but there was no association between social dominance goals and peer nominations for not following the rules for African American students (see Figure 2).

**Intimacy goals.** Intimacy goals were associated only with subsequent self-reports of effort. Thus, striving to have close relationships with friends in the fall was associated with increased levels of effort and involvement in tasks in the classroom in the spring. However, striving for an intimacy goal in the fall had no association with disruptive behavior, peer nominations, or grades.

**Popularity goals.** The relation of popularity goals to subsequent effort, disruptive behavior, and peer nominations of effort/gets good grades varied by gender and ethnicity. Specifically, there was a Popularity Goals  $\times$  Gender  $\times$  Ethnicity interaction indicating that the effects of popularity goals were different for African American girls compared with all other groups. Specifically, for African American girls, the endorsement of popularity goals in the fall was associated with lower subsequent effort, higher subsequent disruptive behavior, and fewer peer nominations of effort/gets good grades, whereas for all other groups, the relation of popularity goals to these outcomes was null (see Figures 3, 4, and 5).

One main effect was not qualified by any interaction: popularity goals were negatively associated with subsequent peer nominations of not following the rules. Thus, students who reported in the

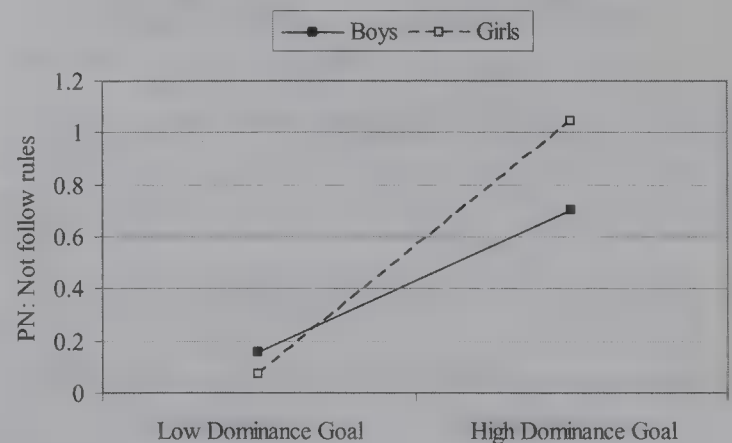


Figure 1. Dominance Goal  $\times$  Gender interaction for peer nominations (PNs) of not following the rules (sixth grade).



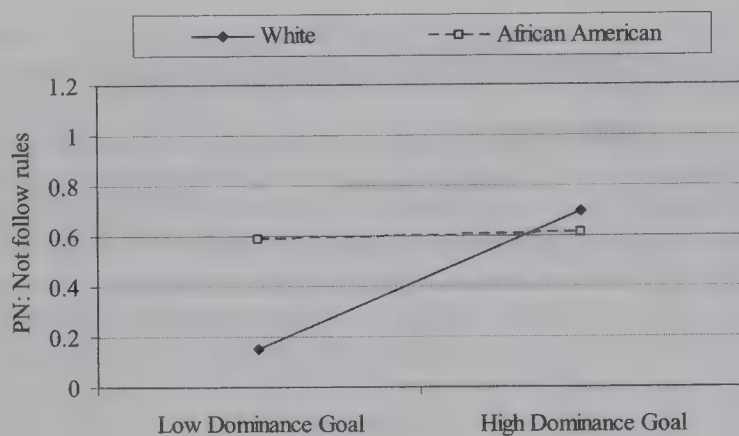


Figure 2. Dominance Goal  $\times$  Ethnicity interaction for peer nominations (PNs) of not following the rules (sixth grade).

fall that they strived to achieve high social status were less likely to get nominations from peers as being students who did not follow the school rules.

#### Multiple Regression Analyses Examining Social Goals, Engagement, and Achievement in Seventh Grade (Middle School)

Next, we conducted separate regression analyses for the five engagement and achievement variables in seventh grade (students' first year in middle school; see Table 4) with the same model-building procedures as described for sixth grade.

**Social dominance goals.** The pattern we found for social dominance goals and academic adjustment in seventh grade was similar to that found for sixth grade. Social dominance goals were positively associated with subsequent disruptive behavior as well as peer nominations for not following the school rules. Additionally, in seventh grade, a negative association was found between social dominance goals and subsequent self-reports of effort. Thus, endorsement of social dominance goals in the beginning of middle school was associated with less positive forms of engagement (i.e., effort) and more negative

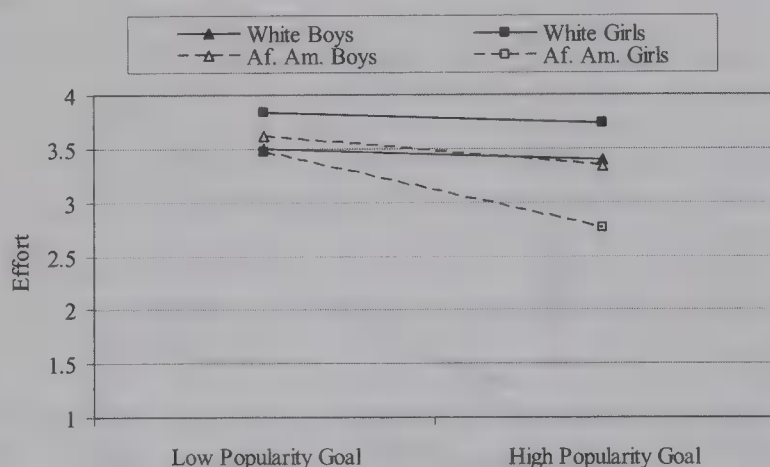


Figure 3. Popularity Goal  $\times$  Gender  $\times$  Ethnicity interaction for effort (sixth grade). Af. Am. = African American.

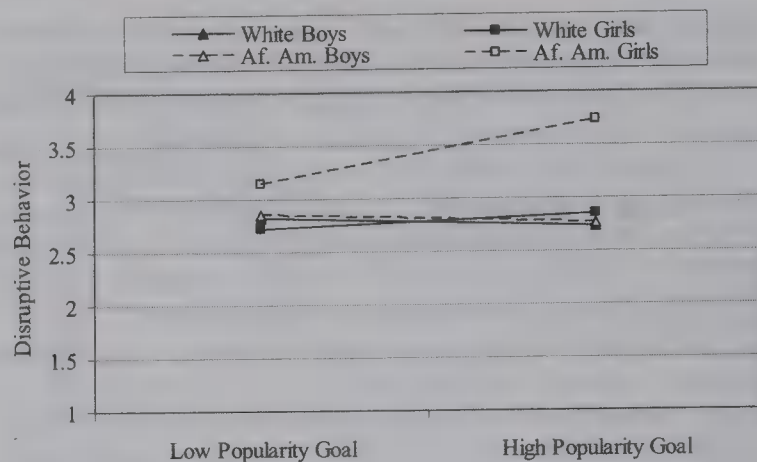


Figure 4. Popularity Goal  $\times$  Gender  $\times$  Ethnicity interaction for disruptive behavior (sixth grade). Af. Am. = African American.

forms of engagement (i.e., disruptive behavior and peer nominations of being a student who did not follow the rules) later in the school year. Similar to the findings in sixth grade, social dominance goals were not associated with peer nominations for effort/gets good grades in seventh grade.

In seventh grade, social dominance goals were negatively associated with subsequent achievement; however, a Social Dominance Goals  $\times$  Ethnicity interaction on achievement indicated that there was a negative association for social dominance goals and subsequent grades for White students, but there was no association for social dominance goals and grades for African American students (see Figure 6).

**Intimacy goals.** In the seventh grade, we found the same positive association between intimacy goals and subsequent effort that we found in the sixth grade. In addition, we found a positive association between intimacy goals and subsequent peer nominations of effort/good grades. Thus, striving to have close relationships with friends in the fall was associated with increased levels of self-reported and peer nominations for effort in the classroom in the spring.

**Popularity goals.** The three-way Popularity Goals  $\times$  Gender  $\times$  Ethnicity interactions that we found for several aspects of

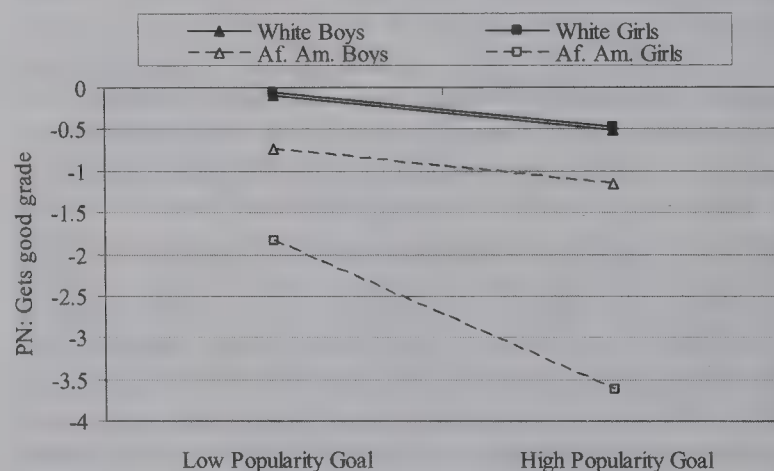


Figure 5. Popularity Goal  $\times$  Gender  $\times$  Ethnicity interaction for peer nominations (PNs) of trying hard and getting good grades (sixth grade). Af. Am. = African American.

Table 4

*Standardized Regression Coefficients for Predicting Academic Adjustment in the Spring From Social Goals in the Fall (Seventh Grade)*

Variable in equation	Adjustment variable				
	Effort ( $\beta$ )	Disruptive ( $\beta$ )	GPA ( $\beta$ )	PN: Effort/grades ( $\beta$ )	PN: Not following rules ( $\beta$ )
Dominance goals	-.15**	.16**	-.37***	-.06	.24**
Popularity goals	-.02	.06	.24**	.11 <sup>†</sup>	-.05
Intimacy goals	.21***	-.03	.08 <sup>†</sup>	.12*	.03
Gender	-.02	-.25***	-.03	.09	-.36***
Ethnicity	.11*	.01	-.44***	-.20*	.04
Gender $\times$ Ethnicity	-.07	.23***	.18*	.01	.07
Dominance $\times$ Ethnicity			.28**		
Popularity $\times$ Ethnicity			-.23**		
$R^2$	.07***	.12***	.20***	.08***	.13***

*Note.* For gender coding, 1 = female; for ethnicity coding, 1 = African American. GPA = grade point average; PN = peer nomination.  
<sup>†</sup>  $p < .06$ . \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

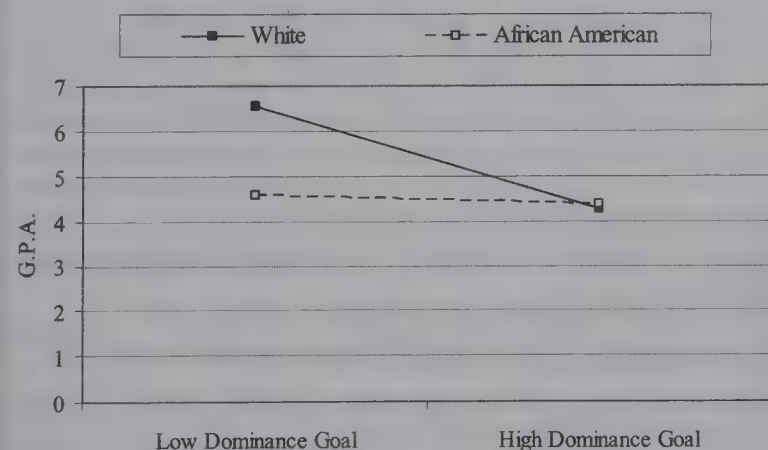
engagement in sixth grade were not found in seventh grade. In seventh grade, popularity goals were associated only with grades, and this association was qualified by a Popularity Goals  $\times$  Ethnicity interaction. The interaction indicated that the endorsement of popularity goals in the beginning of middle school was positively associated with subsequent grades for White students, but there was no association for African American students (see Figure 7).

### Discussion

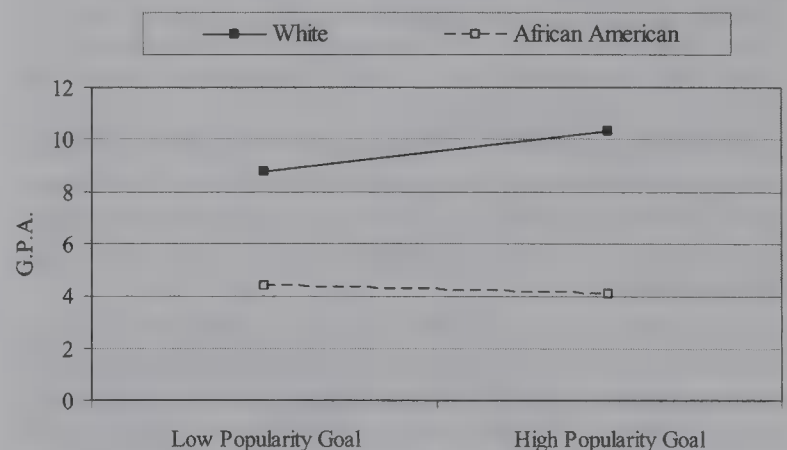
The present research found support for the proposal that social dominance goals would have negative implications for students' engagement and achievement in school. In line with McClelland's (1985) conceptualization, factor analyses supported the assertion that social dominance goals are distinct from popularity and intimacy goals. Social dominance goals have not been the focus of recent research on students' social goals (see Anderman, 1999; Ojanen, Gronroos, & Salmivalli, 2005; Ryan & Shim, 2006a; Wentzel, 1996), but our findings indicate they are important in understanding the implications of social goals for academic adjustment during early adolescence. Dominant and aggressive be-

haviors are prevalent and important to social status in school (see Cillessen & Rose, 2005, for a review). Although there may be some benefits for dominance strivings in terms of social adjustment (as established in previous research, see Hawley et al., 2007), findings of the present study establish that there are serious costs regarding academic adjustment.

In both sixth and seventh grades (students' last year of elementary and first year of middle school), student endorsements of social dominance goals in the fall were associated with maladaptive forms of engagement and lower achievement in the spring. Although effects were not significant for every indicator of engagement, results did encompass both self-reported and peer nominated measures of engagement. When students are focused on establishing power over their peers (making others do what they want and establishing their toughness), they are more likely to act in ways that disrupt classes and are less likely to put effort into their school work. Not surprisingly, social dominance goals were associated with lower achievement. This pattern is not surprising as establishing power over others is inconsistent with the social responsibility, cooperation, and general good citizenship that teachers encourage and expect in the classroom.



*Figure 6.* Dominance Goal  $\times$  Ethnicity interaction for grade point average (G.P.A.; seventh grade).



*Figure 7.* Popularity Goal  $\times$  Ethnicity interaction for grade point average (G.P.A.; seventh grade).



The findings regarding ethnic and gender differences in the level of social goals were in line with McClelland's (1985) theorizing about need for power and Majors's (1991) theorizing about identity development. McClelland theorized that a group's loss of power in society leads to a compensatory increase in the power motive, and Majors theorized that African American males assume a cool pose of toughness, bravado, and emotional detachment in response to discrimination. Congruent with the ideas of both of these scholars, when compared with White students, African American students were higher on social dominance goals in both sixth and seventh grades and lower on intimacy goals in seventh grade. Additionally, being male was associated with increased social dominance and decreased intimacy goals in both sixth and seventh grades. Thus, ethnicity and gender both contributed to increases in social dominance goals. This pattern is consistent with research documenting that these groups are more at risk for disengagement and underachievement in school (Graham & Hudley, 2005). For males, striving for social dominance among peers may be a way of achieving society's ideal masculine traits (e.g., independent, assertive, and powerful; Pleck, Sonenstein, & Ku, 1993). For African American students, striving for social dominance among peers may also be a way to assert oneself and gain control in the school setting in reaction to discrimination (see Simons et al., 2002). Our findings dovetail with research showing that perceptions of discrimination are associated with academic disengagement and disruptive behavior at school (Taylor, Casten, Flickinger, Roberts, & Fulmore, 1994). Collectively, our findings suggest that social dominance goals may be an important aspect of motivation relevant to understanding masculinity, ethnicity, and engagement in school.

There were group differences in the mean levels of social goals, but the pattern of effects for social dominance goals was quite similar across gender and ethnicity. An Ethnicity  $\times$  Social Dominance Goals interaction in sixth grade indicated that, although social dominance goals were positively associated with peer nominations of not following rules for White students, there was no association for African American students. An Ethnicity  $\times$  Social Dominance Goals interaction found in seventh grade indicated that, although social dominance goals were negatively associated with achievement for White students, there was no association for African American students. Thus, there was some suggestion that the maladaptive effects of social dominance goals may not have been quite as pervasive for African American students as for White students. Although social dominance goals may be a negative factor, other issues may be more important to understanding African American students' engagement and achievement. However, even for African American students, the pattern revealed that social dominance goals were maladaptive (e.g., two out of three indicators in sixth grade and three out of four indicators in seventh grade). Further, endorsement of social dominance goals did not promote engagement and achievement in school for any students in either year.

The findings were more limited, but quite consistent, regarding intimacy goals. In both sixth and seventh grades, endorsement of intimacy goals was positively associated with subsequent self-reports of effort. In seventh grade, endorsement of intimacy goals was associated also with peer nominations of effort/good grades. Thus, consistent with prior research, striving for intimacy with peers seems a positive orientation toward peer relations that also

facilitates positive behavior in the classroom (Anderman, 1999; Wentzel, 1993). The desire to make peers happy, understand their feelings, and help them when they need it is consistent with the values teachers would promote among peers in the classroom. The pattern of effects for intimacy goals was identical across gender and ethnicity in both elementary and middle school years.

For the most part, popularity goals were not associated with engagement and achievement in sixth or seventh grade. This pattern, in combination with the limited findings for intimacy goals, supported our proposal that social dominance goals are an important but overlooked aspect of social goals, because social dominance goals were the strongest predictor of academic adjustment. However, there was an exception in the effects of popularity goals for African American girls. For African American girls in sixth grade, popularity goals were associated with maladaptive engagement (i.e., low effort, high disruptive behavior, and low peer nominations as a student who tries hard and gets good grades). Thus, the endorsement of popularity goals (wanting to be popular and get everyone to like them the best) seems to have a different meaning and implications for African American girls. African American girls may pursue and maintain popularity in a way that is distinct from that of other students. Cultural differences in the female role between African American and White students may partially explain how and why strivings for social status play out so differently for African American girls in the school setting. Scholars have noted that there is an emphasis on being strong and outspoken (in addition to caring and nurturing) for girls in African American compared with European American cultures (Basow & Rubin, 1999; Binion, 1990). African American girls who are popular (or trying to be popular) may have a strong vocal presence that does not conform to teachers' expectations of a good girl (i.e., quiet and obedient) in the classroom. Thus, teachers may respond differently to popular girls depending on their ethnicity (see Wright, 2005, for a qualitative study documenting such dynamics). It is interesting that in middle school, this pattern of academic maladjustment ensuing from popularity goals was not found. Instead, popularity goals were associated with higher grades for White students but were not associated with grades for African American students. Further investigation of the nature and meaning of popularity for boys and girls from diverse ethnic groups with attention to academics and teacher perceptions and behaviors could shed some light on how social and academic issues may intersect differently depending on ethnicity and gender.

Future research that uses additional methodologies could be informative. For example, student reports about social goals could be considered with observational data of regular classroom activities or specific tasks assigned to students with peers to provide insight regarding how social goals relate to what students say or do in the classroom or in social tasks. Such research could complement the broader patterns documented in the present study with a richer description of discourse and specific behaviors for both students and teachers.

Another important direction for future research is to investigate school and classroom factors that may minimize the pursuit of social dominance goals or ameliorate their impact on engagement and achievement. If the power motive is a basic human motive, can teachers help channel this motive toward more positive types of social goals, such as leadership? McClelland (1985) believed individuals' motives could be shaped and altered not only to improve



individual outcomes but to better society at large. Examinations of how the school and peer group contexts foster social dominance goals could provide insight about how to structure settings to better support students' learning and achievement.

Although the present findings raise many questions and interesting avenues for future research, the pattern of results regarding social dominance goals and academic adjustment is quite clear. In both sixth and seventh grades, for all students, pursuit of social dominance was associated with maladaptive forms of engagement and lower achievement. These findings contribute to theory on social motivation by highlighting a social goal that has received little attention to date. Further, these findings broaden the understanding of the factors that contribute to engagement and achievement in school with attention to gender and ethnicity. In conclusion, the present research highlights that attention to both social and academic competencies of students, as well as how they intersect, is important for a full appreciation of young adolescents' success in school.

## References

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*, 261–271.
- Anderman, L. H. (1999). Classroom goal orientation, school belonging and social goals as predictors of students' positive and negative affect following the transition to middle school. *Journal of Research & Development in Education, 32*, 89–103.
- Basow, S. A., & Rubin, L. R. (1999). Gender influences on adolescent development. In N. G. Johnson, M. C. Roberts, & J. Worell (Eds.), *Beyond appearance: A new look at adolescent girls* (pp. 25–52). Washington, DC: American Psychological Association.
- Binion, V. J. (1990). Psychological androgyny: A Black female perspective. *Sex Roles, 22*, 487–507.
- Bukowski, W. M., & Hoza, B. (1989). Popularity and friendship: Issues in theory, measurement, and outcome. In T. J. Berndt & G. W. Ladd (Eds.), *Peer relationships in child development* (pp. 15–45). New York: Wiley.
- Cillessen, A. H., & Mayeux, L. (2004). Sociometric status and peer group behavior: Previous findings and current directions. In J. B. Kupersmidt & K. A. Dodge (Eds.), *Children's peer relations* (pp. 3–20). Washington, DC: American Psychological Association.
- Cillessen, A. H., & Rose, A. J. (2005). Understanding popularity in the peer system. *Current Directions in Psychological Science, 14*, 102–105.
- Cunningham, M., & Meunier, L. N. (2004). The influence of peer experiences on bravado attitudes among African American males. In N. Way & J. W. Chu (Eds.), *Adolescent boys: Exploring diverse cultures of boyhood* (pp. 219–232). New York: New York University Press.
- Dowson, M., & McInerney, D. M. (2001). Psychological parameters of students' social and work-avoidance goals: A qualitative investigation. *Journal of Educational Psychology, 93*, 35–42.
- Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95*, 256–273.
- Eccles, J. (1993). School and family effects on the ontogeny of children's interests, self-perceptions, and activity choices. In J. Jacobs (Ed.), *Nebraska symposium on motivation 1992* (Vol. 40, pp. 145–208). Lincoln: University of Nebraska Press.
- Eccles, J. (2004). Schools, academic motivation, and stage-environment fit. In R. M. Lerner & L. Steinberg (Eds.), *Handbook of adolescent psychology* (2nd ed., pp. 125–153). New York: Wiley.
- Eder, D. (1985). The cycle of popularity: Interpersonal relations among female adolescents. *Sociology of Education, 58*, 154–165.
- Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist, 34*, 169–189.
- Elliot, A. J., & Dweck, C. S. (2005). *Handbook of competence and motivation*. New York: Guilford.
- Erdley, C. A., Cain, K. M., Loomis, C. C., Dumas-Hines, F., & Dweck, C. (1997). Relations among children's social goals, implicit personality theories, and responses to social failure. *Developmental Psychology, 33*, 263–272.
- Ford, M. E. (1992). *Motivating humans: Goals, emotions, and personal agency beliefs*. Newbury Park, CA: Sage.
- Gorman, A. H., Kim, J., & Schimmelbusch, A. (2002). The attributes adolescents associate with peer popularity and teacher preference. *Journal of School Psychology, 40*, 143–165.
- Graham, S., & Hudley, C. (2005). Race and ethnicity in the study of motivation and competence. In A. J. Elliot, & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 392–413). New York: Guilford.
- Graham, S., Taylor, A. Z., & Hudley, C. (1998). Exploring achievement values among ethnic minority early adolescents. *Journal of Educational Psychology, 90*, 606–620.
- Grotpeter, J. K., & Crick, N. R. (1996). Relational aggression, overt aggression, and friendship. *Child Development, 67*, 2328–2338.
- Hawley, P. H. (2003). Prosocial and coercive configurations of resource control in early adolescence: A case for the well-adapted Machiavellian. *Merrill-Palmer Quarterly, 49*, 279–309.
- Hawley, P. H., Little, T. D., & Rodkin, P. C. (2007). *Aggression and adaptation: The bright side to bad behavior*. Hillsdale, NJ: Erlbaum.
- Heinlein, L. M., & Shinn, M. (2000). School mobility and student achievement in an urban setting. *Psychology in the Schools, 37*, 349–357.
- Horst, S. J., Finney, S. J., & Barron, K. E. (2007). Moving beyond academic achievement goal measures: A study of social achievement goals. *Contemporary Educational Psychology, 32*, 667–698.
- Jarvinen, D. W., & Nicholls, J. G. (1996). Adolescents' social goals, beliefs about the causes of social success, and satisfaction in peer relations. *Developmental Psychology, 32*, 435–441.
- Juvonen, J., & Murdock, T. B. (1993). How to promote social approval: Effects of audience and achievement outcome on publicly communicated attributions. *Journal of Educational Psychology, 85*, 365–376.
- Juvonen, J., & Murdock, T. B. (1995). Grade-level differences in the social value of effort: Implications for self-presentation tactics of early adolescents. *Child Development, 66*, 1694–1705.
- Kaplan, A., & Maehr, M. L. (1999). Achievement goals and student well-being. *Contemporary Educational Psychology, 24*, 330–358.
- Ladd, G. W., Kochenderfer, B. J., & Coleman, C. C. (1997). Classroom peer acceptance, friendship, and victimization: Distinct relational systems that contribute uniquely to children's school adjustment? *Child Development, 68*, 1181–1197.
- LaFontana, K. M., & Cillessen, A. H. N. (2002). Children's perceptions of popular and unpopular peers: A multimethod assessment. *Developmental Psychology, 38*, 635–647.
- Leventhal, T., & Brooks-Gunn, J. (2000). The neighborhoods they live in: Effects of neighborhood residence upon child and adolescent outcomes. *Psychological Bulletin, 126*, 309–337.
- Majors, B. (1991). Nonverbal behaviors and communication styles among African Americans. In R. L. Jones (Ed.), *Black psychology* (pp. 269–294). Hampton, VA: Cobb & Henry.
- Majors, B., & Billson, J. M. (1992). *Cool pose: The dilemmas of Black manhood in America*. New York: Macmillan.
- McAdams, D. P. (1988). *Power, intimacy, and the life story: Personological inquiries into identity*. Chicago: Dorsey Press.
- McClelland, D. C. (1985). *Human motivation*. New York: Cambridge University Press.
- Merten, D. E. (1997). The meaning of meanness: Popularity, competition, and conflict among junior high school girls. *Sociology of Education, 70*, 175–191.



- Midgley, C., Kaplan, A., Middleton, M., Maehr, M. L., Urdan, T., Anderman, L. H., et al. (1998). The development and validation of scales assessing students' achievement goal orientations. *Contemporary Educational Psychology*, 23, 113-131.
- Miserandino, M. (1996). Children who do well in school: Individual differences in perceived competence and autonomy in above-average children. *Journal of Educational Psychology*, 88, 203-214.
- Murray, H. A. (1938). *Explorations in personality*. New York: Oxford University Press.
- Nicholls, J. G., Cheung, P. C., Lauer, J., & Pataschnick, M. (1989). Individual differences in academic motivation: Perceived ability, goals, beliefs and values. *Learning and Individual Differences*, 1, 63-84.
- Ojanen, T., Gronroos, M., & Salmivalli, C. (2005). An interpersonal circumplex model of children's social goals: Links with peer-reported behavior and sociometric status. *Developmental Psychology*, 41, 699-710.
- Patrick, H. (1997). Social self-regulation: Exploring the relations between children's social relationships, academic self-regulation, and school performance. *Educational Psychologist*, 32, 209-220.
- Patrick, H., Turner, J. C., Meyer, D. K., & Midgley, C. (2003). How teachers establish psychological environments during the first days of school: Associations with avoidance in mathematics. *Teachers College Record*, 105, 1521-1558.
- Pellegrini, A. D., & Bartini, M. (2001). Dominance in early adolescent boys: Affiliative and aggressive dimensions and possible functions. *Merrill-Palmer Quarterly*, 47, 142-163.
- Pintrich, P. R. (2000). An achievement goal theory perspective on issues in motivation terminology, theory, and research. *Contemporary Educational Psychology*, 25, 92-104.
- Pleck, J. H., Sonenstein, F. L., & Ku, L. C. (1993). Masculinity ideology and its correlates. In S. Oskamp & M. Constanzo (Eds.), *Gender issues in contemporary society* (pp. 85-110). Newbury Park: Sage.
- Rodkin, P. C., Farmer, T. W., Pearl, R., & Van Acker, R. (2000). Heterogeneity of popular boys: Antisocial and prosocial configurations. *Developmental Psychology*, 36, 14-24.
- Rose, A. J., & Rudolph, K. D. (2006). A review of sex differences in peer relationship processes: Potential trade-offs for the emotional and behavioral development of girls and boys. *Psychological Bulletin*, 132, 98-131.
- Rose, A. J., Swenson, L. P., & Waller, E. M. (2004). Overt and relational aggression and perceived popularity: Developmental differences in concurrent and prospective relations. *Developmental Psychology*, 40, 378-387.
- Rowe, D., Vazsonyi, A., & Flannery, D. (1994). No more than skin deep: Ethnic and racial similarity in developmental processes. *Psychological Review*, 101, 396-413.
- Rubin, K. H., Bukowski, W., & Parker, J. G. (2006). Peer interactions, relationships, and groups. In W. Damon, & R. Lerner (Series Eds.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (6th ed., pp. 571-645). New York: Wiley.
- Ruble, D. N., Martin, C. L., & Berenbaum, S. A. (2006). Gender development. In N. Eisenberg (Ed.), *Handbook of child psychology: Social, emotional, and personality development* (pp. 858-932). New York: Wiley.
- Ryan, A. M. (2001). The peer group as a context for the development of young adolescent motivation and achievement. *Child Development*, 72, 1135-1150.
- Ryan, A. M., Hicks, L., & Midgley, C. (1997). Social goals, academic goals, and avoiding seeking help in the classroom. *Journal of Early Adolescence*, 17, 152-171.
- Ryan, A. M., Kiefer, S. M., & Hopkins, N. B. (2004). Young adolescents' social motivation: An achievement goal perspective. In M. L. Maehr & P. R. Pintrich (Eds.), *Advances in motivation and achievement* (Vol. 13, pp. 301-330). Amsterdam, the Netherlands: Elsevier.
- Ryan, A. M., & Shim, S. (2006a). *An exploration of young adolescents' social achievement goals and social adjustment in middle school*. Manuscript submitted for publication.
- Ryan, A. M., & Shim, S. (2006b). Social achievement goals: The nature and consequences of different orientations toward social competence. *Personality and Social Psychology Bulletin*, 32, 1246-1263.
- Savin-Williams, R. C., & Berndt, T. J. (1990). Friendship and peer relations. In S. S. Feldman & G. R. Elliott (Eds.), *At the threshold: The developing adolescent* (pp. 277-307). Cambridge, MA: Harvard University Press.
- Simons, R., Murry, V., McLoyd, V., Lin, K., Cutrona, C., & Conger, R. (2002). Discrimination, crime, ethnic identity and parenting as correlates of depressive symptoms among African American children: A multilevel analysis. *Development and Psychopathology*, 14, 371-393.
- Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology*, 85, 571-581.
- Stevenson, H. C. (2004). Boys in men's clothing: Racial socialization and neighborhood safety as buffers to hypervulnerability in African American adolescent males. In N. Way & J. Y. Chu (Eds.), *Adolescent boys: Exploring diverse cultures of boyhood* (pp. 59-77). New York: New York University Press.
- Taylor, R., Casten, R., Flickinger, S., Roberts, D., & Fulmore, C. (1994). Explaining the school performance of African-American adolescents. *Journal of Research on Adolescence*, 4, 21-44.
- Urdan, T. (1997). Achievement goal theory: Past results, future directions. In M. Maehr & P. R. Pintrich (Eds.), *Advances in motivation and achievement* (Vol. 10, pp. 99-141). Greenwich, CT: JAI Press.
- Urdan, T., & Maehr, M. (1995). Beyond a two-goal theory of motivation: A case for social goals. *Review of Educational Research*, 65, 213-244.
- Veroff, J. (1957). Development and validation of a projective measure of power motivation. *Journal of Abnormal and Social Psychology*, 54, 1-8.
- Veroff, J., Depner, C., Kulka, R., & Douvan, E. (1980). Comparison of American motives: 1957 versus 1976. *Journal of Personality and Social Psychology*, 39, 1249-1262.
- Wentzel, K. R. (1992). Motivation and achievement in adolescence. A multiple goals perspective. In D. Schunk & J. Meece (Eds.), *Student perceptions in the classroom: Causes and consequences* (pp. 287-306). Hillsdale, NJ: Erlbaum.
- Wentzel, K. R. (1993). Motivation and achievement in early adolescence: The role of multiple classroom goals. *Journal of Early Adolescence*, 13, 4-20.
- Wentzel, K. R. (1996). Social and academic motivation in middle school: Concurrent and long-term relations to academic effort. *Journal of Early Adolescence*, 16, 390-406.
- Wentzel, K. R. (2000). What is it that I'm trying to achieve? Classroom goals from a content perspective. *Contemporary Educational Psychology*, 25, 105-115.
- Wentzel, K. R. (2005). Peer relationships, motivation, and academic performance at school. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 279-296). New York: Guilford.
- Wigfield, A., Eccles, J. S., Schiefele, U., Rosser, R. W., & Davis-Kean, P. (2006). Development of achievement motivation. In W. Damon & R. M. Lerner (Eds.), *Handbook of child psychology* (6th ed., pp. 933-1002). New York: Wiley.
- Winter, D. G. (1973). *The power motive*. New York: Free Press.
- Winter, D. G., & Stewart, A. J. (1978). Power motivation. In H. London & J. Exner (Eds.), *Dimensions of personality* (pp. 391-447). New York: Wiley.
- Wright, C. (2005). Black femininities go to school: How young Black females navigate race and gender. In G. Lloyd (Ed.), *Problem girls: Understanding and supporting troubled and troublesome girls and young women* (pp. 103-113). New York: Routledge Falmer.

Received October 31, 2006

Revision received July 3, 2007

Accepted August 4, 2007 ■

# Do Peers Contribute to the Likelihood of Secondary School Graduation Among Disadvantaged Boys?

Marie-Hélène Véronneau, Frank Vitaro, Sara Pedersen, and Richard E. Tremblay  
Université de Montréal

This 17-year longitudinal study tested whether low peer-perceived acceptance and association with aggressive-disruptive friends during preadolescence predicted students' failure to graduate from secondary school. Participants were 997 Caucasian, French-speaking boys from low-socioeconomic status, urban neighborhoods. The boys were recruited in kindergarten (age 6) and followed through early adulthood (age 23). Low levels of prosocial behaviors and high levels of aggressive-disruptive behaviors in childhood were expected to predict negative preadolescent peer experiences. Adolescent academic achievement and school commitment were expected to mediate the link between preadolescent peer experiences and early adulthood graduation status. Results of structural equation modeling analyses tended to support these hypotheses. Greater childhood aggression-disruptiveness positively predicted friends' preadolescent aggression-disruptiveness. Having aggressive-disruptive friends, in turn, was related to a lower likelihood of graduation. Lower academic achievement and school commitment partially mediated the association between friend characteristics and graduation. Peer acceptance did not contribute to graduation.

**Keywords:** friendship, peer acceptance, academic adjustment, school graduation, longitudinal studies

Failure to graduate from secondary school has been related to problems at both individual and societal levels. At the individual level, such psychosocial problems as difficulties finding and maintaining employment, loss of wages, psychological maladjustment, and lower rates of community involvement have been associated with low educational attainment (Kaplan, Damphousse, & Kaplan, 1994; Kerckhoff & Bell, 1998; McCaul, Donaldson, Coladarsi, & Davis, 1992). At the societal level, major social and economic changes, such as the aging of the workforce, technological innovations, and market globalization (Organisation for Economic Co-Operation and Development, 2005) have put increased pressure on industrialized countries to maintain a well-trained workforce. A high rate of secondary school graduation is one important goal societies must achieve in order to meet this global educational challenge.

Research that enhances our understanding of the experiences that might influence students' abilities to earn secondary school

diplomas is an essential first step toward increasing the secondary school graduation rate. The role of familial and individual factors in the process leading to secondary school graduation was extensively investigated during the second half of the 20th century (see reviews by Ekstrom, Goertz, Pollack, & Rock, 1986, and Rumberger, 1987). In contrast, studies assessing the contribution of peers to secondary school graduation are still scarce, despite the important role that peers play in adolescents' lives (Hymel, Comfort, Schonert-Reichl, & McDougall, 1996; Vitaro, Larocque, Janosz, & Tremblay, 2001). The current study focuses on peer-related variables as potentially important contributors to individual graduation outcomes in boys. Boys graduate from secondary school at much lower rates than girls, and thus are of particular concern (Bowlby & McMullen, 2005).

## Preliminary Conceptual Clarifications

As put forward by Christenson, Sinclair, Lehr, and Godber (2001), graduation (or the failure thereof) from school should be distinguished from school persistence and its opposite, school interruption (i.e., school dropout). Graduation refers to the completion of the school program as attested to by a diploma, whereas school dropout refers to an interruption of school attendance, which may be definitive (i.e., permanent dropout) or followed by a return to school (i.e., temporary dropout). Christenson and her colleagues found this distinction useful for designing interventions aimed at enhancing students' academic attainment; we suggest that this distinction is also crucial in the context of nonexperimental, descriptive research. In fact, Entwisle, Alexander, and Olson (2004) found that temporary dropouts who eventually graduate usually become well-adjusted adults, whereas permanent dropouts (i.e., those who do not graduate) often suffer psychosocial problems. In other words, it is not the interruption of school attendance

---

Marie-Hélène Véronneau and Richard E. Tremblay, Department of Psychology and Research Unit on Children's Psychosocial Maladjustment, Université de Montréal, Montreal, Quebec, Canada; Frank Vitaro, School of Psychoeducation and Research Unit on Children's Psychosocial Maladjustment, Université de Montréal; Sara Pedersen, Research Unit on Children's Psychosocial Maladjustment, Université de Montréal.

This manuscript was part of Marie-Hélène Véronneau's doctoral dissertation. Its preparation was made possible by a doctoral fellowship awarded to Marie-Hélène Véronneau by the Social Sciences and Humanities Research Council of Canada.

Correspondence concerning this article should be addressed to Marie-Hélène Véronneau, Research Unit on Children's Psychosocial Maladjustment, Université de Montréal, C.P. 6128, Succursale Centre-Ville, Montreal, Quebec H3C 3J7, Canada. Email: mh.veronneau-mcardle@umontreal.ca



per se, but rather the failure to complete the full academic program that most likely affects adjustment.

Nevertheless, because few studies use graduation as an outcome, it is useful to rely on dropout research to help identify relevant risk factors for the failure to graduate. In a review of the dropout literature, Rumberger (1987) made a distinction between structural risk factors for dropout (e.g., socioeconomic status [SES] and gender) and manipulable variables that are likely to explain the process of school disengagement (e.g., students' behaviors or some aspects of the social milieu). Although Rumberger acknowledged the predictive value of the structural risk factors, he suggested that future research should focus on manipulable variables because these can be targeted by interventions aimed at increasing graduation rates. This is why we focused on potentially manipulable social processes involving peers.

### An Integrative Theoretical Model of Secondary School Graduation

The current research was guided by an integrative model of secondary school graduation. This global model includes different types of predictors and processes that have been proposed by two complementary theoretical perspectives, namely, the social interactional perspective and the participation-identification perspective of school withdrawal.

#### *Sociofamilial, Behavioral, and Peer-Related Antecedents of the Failure to Graduate: Contributions of the Social Interactional Perspective*

The social interactional perspective was introduced by Patterson and his colleagues to study the pathways toward antisocial behavior, a broad concept that includes school dropout as well as many other socially deviant behaviors (Patterson, DeBaryshe, & Ramsey, 1989; Patterson, Forgatch, Yoerger, & Stoolmiller, 1998; Patterson, Reid, & Dishion, 1992). From this perspective, the failure to graduate is the end result of a developmental pathway that starts at home, where young boys develop a coercive behavior pattern. Such behavior then leads to rejection by normative in-school peers and to the affiliation with deviant peers who reinforce problem behaviors and attitudes that are incompatible with academic success.

More precisely, according to this view, the family context in the preschool years—especially inefficient, harsh, and inconsistent parenting practices—often paves the way for antisocial behavior. Consistent with Rumberger's (1987) review of risk factors for school dropout, children born to families of lower SES are particularly likely to be raised in a risky family climate. The high stress levels experienced by parents living in precarious financial conditions in combination with their typically lower educational attainment may heighten the likelihood that they will use coercive and inconsistent parenting practices. In addition, low-SES parents are less likely to adopt the beliefs and behaviors that contribute positively to their children's academic success (Davis-Kean, 2005).

When a child raised in a risky family context enters school, the coercive pattern of behavior he or she has developed in response to the family environment may interfere with adaptation to the new milieu, because such behaviors are in contradiction to the prevailing social norms for effective interpersonal interactions outside of

the family environment. Among other aspects of the child's psychosocial adjustment, experiences with peers may be severely affected by this lack of social skills. In the context of the social interactional perspective, two types of peer experiences have been examined, often separately.

First, peer acceptance or rejection by the peer group is thought to be crucial to children's psychosocial development. In Sullivan's (1953) interpersonal theory, peers act as socializing agents, rewarding acceptable behaviors in other children through positive interactions. These positive interactions lead to the successful fulfillment of the child's need for group belongingness. Conversely, peers discourage unacceptable conduct by avoiding interactions with—or rejecting—children who display aggressive, disruptive, or coercive behaviors. Empirically, several studies have established that children displaying high levels of aggressive behavior and low levels of prosocial behavior are at greater risk of being rejected by their peers (Coie, Dodge, & Kupersmidt, 1990; Ladd & Troop-Gordon, 2003; Taylor, 1989; Vitaro, Tremblay, Gagnon, & Boivin, 1992; Wentzel, 2003). In turn, low levels of prosocial behaviors, high levels of aggressive behavior, and peer rejection are related to a heightened risk of school dropout (Ollendick, Weist, Borden, & Greene, 1992; Vitaro, Brendgen, Larose, & Tremblay, 2005).

Second, in addition to peer rejection, friends' deviancy also represents a fundamental aspect of the peer experience that is related to school dropout in boys (Vitaro et al., 2001). According to the social interactional perspective, forming friendships with deviant peers is an important step in the pathway toward antisocial behavior and, in the context of the current study, toward the failure to graduate from secondary school. Patterson et al. (1992) theorized that youngsters who are rejected by the peer group affiliate with other aggressive, rejected children in order to fulfill their socioaffective needs, and empirical studies have supported this idea (Brendgen, Vitaro, & Bukowski, 1998; Dishion, Patterson, Stoolmiller, & Skinner, 1991; Laird, Jordan, Dodge, Pettit, & Bates, 2001; Vitaro, Pedersen, & Brendgen, 2007). However, friendships among aggressive and rejected students are usually short-lived and low in quality (Bagwell & Coie, 2004; Dishion, Andrews, & Crosby, 1995; Parker & Asher, 1993). As a result, these unsatisfying friendships may have a negative rather than a positive impact on psychosocial and academic outcomes in children and adolescents (Burk & Laursen, 2005; Ladd, Kochenderfer, & Coleman, 1996; Lansford, Criss, Pettit, Dodge, & Bates, 2003).

This may be especially true if friends of future school dropouts have similarly negative school-related experiences, including academic failure and rejection by peers and school staff, as well as feelings of alienation from school (Ekstrom et al., 1986). Through the process of peer socialization (Kandel, 1978), deviant friends may reinforce or model attitudes and behaviors that are incompatible with school success (Dishion, McCord, & Poulin, 1999). There is preliminary support for the hypothesis that students who associate with friends who reject school are more likely to disengage from school (Battin-Pearson et al., 2000; Pittman, 1991). Such disengagement can lead to truancy and school dropout.

The main strength of the social interactional perspective is its explicit acknowledgment that several types of risk factors work together from the earliest years of life to deter some individuals from following adaptive psychosocial and educational trajectories, such as those leading toward secondary school graduation. This

theory, however, was developed to explain boys' antisocial behavior in general, not the failure to graduate from secondary school in particular. As a result, it is not clear whether the peer processes described earlier (i.e., peer acceptance at the group level and friends' characteristics at the dyadic level) play an active (i.e., mediating) role in the process leading disruptive and socially unskilled boys to encounter school problems and, ultimately, to fail to obtain a secondary school diploma. In addition, the individual psychological processes that might explain or mediate the hypothesized role of peer experiences with regard to school withdrawal are not clearly defined. The participation-identification perspective described by Finn (1989) specifies which psychological processes may be involved and thus complements the social interactional perspective in explaining how negative peer experiences may contribute specifically to the failure to graduate from secondary school.

### *Psychological and Motivational Antecedents of the Failure to Graduate: Contributions of the Participation-Identification Perspective*

Finn (1989) suggested that students undergo a cycle of academic participation and identification with school during their school years. For most students, active participation in school activities leads to positive academic outcomes (e.g., high academic achievement), which in turn reinforce psychological and emotional identification with school. School identification—described as a feeling of belongingness to the school milieu and as the internalization of the school's goals and values—is thought to be essential to academic perseverance and secondary school graduation.

However, for some students, the participation-identification cycle is disrupted, thereby affecting both academic performance and commitment to schooling. These students may fail to internalize the school's goals and values, increasing the odds of premature school withdrawal and failure to graduate. Negative peer relationships are a potential disrupter of this cycle. For example, ostracism by the peer group may generate aversion for classmates and school

in general (see review by Hymel et al., 1996). Having friends with deviant attitudes and behaviors may also influence boys to behave in deviant ways (Dishion et al., 1995), so that school disengagement might be contagious within deviant peer groups.

### *The Impact of Peer Experiences on Secondary School Graduation*

By combining the variables that were identified as crucial elements in the pathway leading to the failure to graduate from secondary school according to either the social interactional or the participation-identification perspectives, we have developed the comprehensive model presented in Figure 1. As suggested by these theoretical perspectives, peer experiences are depicted as central elements in the developmental pathway leading to disengagement from school and, ultimately, the failure to graduate. Still, much empirical research remains to be done in order to confirm the validity of this model.

One important issue that needs to be addressed is the distinct contributions of different aspects of the peer experience (Furman & Robbins, 1985; Hartup, 1996; Sullivan, 1953). Gifford-Smith and Brownell (2003) suggested that most studies of peer relations have examined a single aspect of this experience (e.g., either peer acceptance or friends' characteristics). However, the only way to uncover the unique effects of different types of peer experiences is to incorporate all of these variables within the same study.

Another issue that deserves special attention has been raised by several authors (e.g., Finn, 1989; Hymel et al., 1996; Rumberger, 1987), who have contended that, even though the correlation between early peer difficulties and school dropout is well established, very little is known about the psychological processes behind this relation. In fact, longitudinal studies often focus on observable (social, behavioral, and academic) variables rather than on the psychological process of disengagement.

In order to address these issues, the current study assessed two distinct types of peer experiences that are likely to contribute to secondary school graduation in boys. These include peer accep-

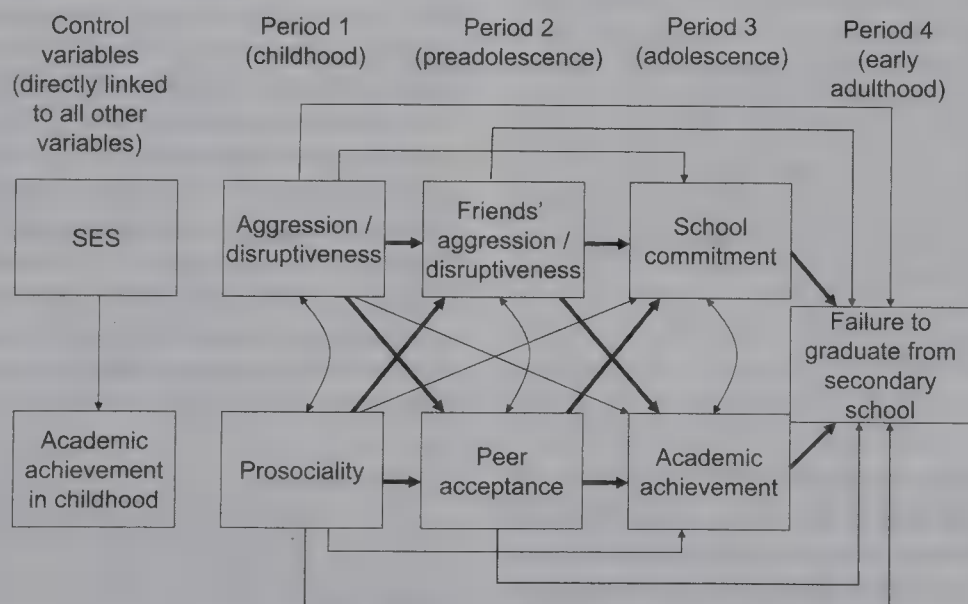


Figure 1. Initial model. Arrows in bold represent the hypothesized indirect (mediation) paths. SES = socioeconomic status.



tance and friends' aggressive behaviors. We also took into account the roles of school commitment and academic achievement as "psychological consequences" of negative peer experiences and, consequently, as proximal precursors of the failure to graduate that could mediate the effects of negative peer experiences (see Figure 1). Other established precursors of graduation were also taken into account as control variables (i.e., SES and early academic achievement).

### Current Study

The longitudinal data used to test the proposed developmental model were collected over several periods of development, including childhood (Period 1), preadolescence (Period 2), adolescence (Period 3), and early adulthood (Period 4).<sup>1</sup>

In Period 1, children's behaviors serve as antecedents of the peer experiences. Specifically, high levels of aggressive-disruptive behaviors as well as low levels of prosocial behaviors are the hypothesized antecedents of low peer acceptance and association with aggressive-disruptive friends in preadolescence (Period 2).

Next, negative preadolescent peer experiences are hypothesized to predict the failure to graduate because they interfere with normal learning activities and contribute to psychological disengagement from school. In other words, peer experiences in preadolescence (Period 2) should predict students' academic achievement and commitment to educational goals and values during adolescence (Period 3). Finally, lower academic achievement and school commitment during adolescence (Period 3) are hypothesized to predict the failure to graduate by early adulthood (Period 4) directly (and act as mediators of the association between negative peer experiences and the failure to graduate).

Lower family SES and academic achievement in childhood are early risk factors for deviant school pathways and, as such, they are used as control variables. In order to perform a more stringent test of mediation, we assumed control variables were related not only to the dependent variable, but also to all other variables in the model.

This study was performed with a community-based sample. However, participants had a relatively high risk of failing to graduate from secondary school, as they were boys recruited from public schools located in low-SES, urban neighborhoods. Indeed, whereas the overall provincial graduation rate for male and female students aged 20 years or less was 71.7% (Education Department, 2001), the rate of graduation among our participants by age 23 was only 53.3%. More precisely, 49.7% of participants graduated without interrupting their schooling, and 3.6% graduated in spite of having temporarily dropped out of school.

### Method

#### Participants

Participants were recruited when they were 6 years old at the end of the 1983–1984 school year, in the context of the Montreal Longitudinal Experimental Study, a larger project aimed at studying the psychosocial development of low-SES, urban boys. This community-based sample initially targeted all boys attending kindergarten in the 53 public schools located in low-SES neighborhoods that were part of the Montreal French school board (Quebec,

Canada). From the 1,161 boys recruited (85.0% of the targeted sample), only the 1,037 boys whose parents were born in Canada and spoke French as a first language were retained in the study. Data on the dependent variable, secondary school graduation by 23 years of age (2001), were available for 98.2% of this sample (1,018 participants). Sixteen participants had died by that time, and 3 had moved out of the country. Because of missing data, the final sample used for the main analysis was composed of 997 participants. (Details on attrition are provided in the Results section.)

#### Measures

All measures were administered in French. Instruments that were only available in English were translated into French and back-translated into English. The back-translations were compared to the original versions of the questionnaires by English-speaking judges who verified that the original meaning was preserved. The timing of assessments and descriptive statistics for all instruments are presented in Table 1. Because the skewness statistic for all variables ranged from  $-0.33$  to  $0.92$ , and kurtosis ranged from  $-1.98$  to  $0.21$ , normality of the data was deemed satisfactory for further analyses.

**SES.** Parents' occupational prestige was computed from their earning and educational levels (Blishen, Carroll, & Moore, 1987). The measure was taken when participants were 6 years old (the 1st year of the study) or at a subsequent year if earning and educational data were not available at the first wave of data collection (see the *Attrition* subsection under the Results section for more details on this imputation procedure). For most participants, a score was available for both parents, so an average score was used.

Blishen et al. (1987) did not mention a specific range of scores for low-SES occupations, but they reported a mean score (42.74) and the standard deviation (13.28) for their instrument. The minimum score on this scale is 17.81, which corresponds to the occupations of newspaper carrier and vendor. The average score for this sample ( $M = 38.45$ ,  $SD = 10.70$ ) is about one-third of a standard deviation below the mean proposed by Blishen et al. It is lower than the mean obtained by a representative sample of boys living in the province of Quebec over the same period ( $M = 42.08$ ,  $SD = 12.09$ ). Examples of occupations that correspond to the average SES score for the current sample are orderlies, mail and postal clerks, and machine tool operators.

**Academic achievement in early elementary school years.** Elementary school teachers reported on students' academic achievement in French (first language) and mathematics in 1985 and 1988 (when the students were ages 7 and 10), which correspond to Grades 1 and 4 for students in age-appropriate classrooms. In order to increase the reliability of the scores obtained from this measure, an average score of academic achievement over the 2 years was computed to represent academic achievement in childhood. For the participants with missing data at one of these assessments (28.3%

<sup>1</sup> Several participants were retained in lower grades: The rate of retention was 10.9% by the end of Grade 1 and tended to increase in subsequent years. Therefore, to be consistent with the developmental periods being studied, data were gathered for each participant at every year, even for those who were in lower grades. Nevertheless, for the sake of simplicity, only the school grades for participants in age-appropriate classrooms are reported in the *Measures* section.

Table 1  
*Age(s) at Assessment, Indices of Central Tendency, and Distributional Properties of Study Measures*

Measure	Age(s) at assessment (years) <sup>a</sup>	N	M	SD	Skewness	Kurtosis
Control variables						
SES	6	997	38.45	10.70	0.82	0.21
Academic achievement	7, 10	934	2.03	0.87	-0.33	-0.41
Period 1						
Aggression-disruptiveness	6, 10	997	6.49	5.51	0.80	-0.17
Prosociality	6, 10	997	7.57	3.90	0.23	-0.48
Period 2						
Friends' aggression-disruptiveness	11, 12	721	-0.06	0.74	0.59	0.05
Peer acceptance	11, 12	924	-0.11	1.58	-0.18	-0.74
Period 3						
School commitment	14-17	883	17.17	3.25	-0.06	-0.02
Academic achievement	14-17	876	1.68	0.85	0.00	-0.25
Period 4						
Failure to graduate	23	981	0.54	0.50	-0.15	-1.98

Note. SES = socioeconomic status.

<sup>a</sup>Where more than one age is listed, the mean score over several assessments was used.

of the final sample), we used the value obtained at the other assessment so they did not have to be excluded from the analyses due to missing data. Final scores ranged on a continuous scale from 0 (*academic failure*) to 4 (*excellent academic performance*). A correlation of .53 ( $p < .001$ ) was found for academic achievement in Grades 1 and 4 and was based on participants with complete data at both times of measurement (64.2% of the total sample).

*Participants' behavior in early elementary school years.* Kindergarten and elementary school teachers reported on students' aggressive-disruptive behaviors and on their prosocial behaviors in 1984 and 1988 (when the students were ages 6 and 10), which correspond to kindergarten and Grade 4 for students in age-appropriate classrooms, using the Social Behavior Questionnaire (Tremblay, Desmarais-Gervais, Gagnon, & Charlebois, 1987; Tremblay, Vitaro, Gagnon, Piché, & Royer, 1992). Aggression-disruptiveness was assessed with items such as "This child bullies," "This child kicks, bites or hits," and "This child is restless and cannot stand still," and prosociality was assessed with items like "This child praises other students" and "This student helps clear up a mess," which were rated 0 (*never*), 1 (*sometimes*), or 2 (*often*). Scores on the 13-item aggression-disruptiveness scale ranged from 0 to 26, with higher scores representing frequent occurrences of aggressive-disruptive behaviors. Cronbach's alphas for this scale were .93 for the kindergarten assessment and .92 for the Grade 4 assessment. Scores on the 10-item prosocial scale ranged from 0 to 20, with higher scores representing frequent occurrences of prosocial behaviors. Cronbach's alphas for this scale were .92 for the kindergarten assessment and .91 for the Grade 4 assessment. Correlations of .47 ( $p < .001$ ) and .23 ( $p < .001$ ) were found between kindergarten and Grade 4 scores for aggressive-disruptive and prosocial behaviors, respectively, and were based on participants with complete data at both times of measurement (94.2% and 94.0% of the total sample, respectively). In order to increase the reliability of the scores obtained from these measures, the average score over the 2 years was used for each scale. For participants with missing data at one of these assessments, we used the value obtained at the other assessment.

*Peer experiences in late elementary school years.* Measures of peer experiences were based on scores obtained in 1989 and 1990 (when the students were ages 11 and 12), which correspond to Grades 5 and 6 for children in age-appropriate classrooms. Again, these measures were averaged over the 2 years.

For peer experiences at the group level, a measure of peer-perceived acceptance in the peer group was used. Scores on two items, namely, "Those who are liked by everyone" and "Those who have very few friends," were combined to create this variable. These items were part of the Pupil Evaluation Inventory (PEI; Pekarik, Prinz, Liebert, Weintraub, & Neale, 1976) that was administered to all children in the classrooms attended by the participants. For each item, students could nominate up to four classmates. The total number of nominations received by each participant on every item was standardized within their classroom. A correlation of .65 ( $p < .001$ ) was found for peer acceptance in Grades 5 and 6 and was based on participants with complete data at both times of measurement (74.9% of the total sample).

For measures of dyadic peer experiences, participants' friends were identified using reciprocated nominations on the item "Those who are your best friends" from the PEI questionnaire. Then, friends' scores on the 20-item aggression-disruptiveness scale of the PEI (including items like, "Those who start a fight over nothing" and "Those who disturb other students who are trying to work") were used to assess their level of aggressive-disruptive behaviors. Cronbach's alphas for this scale were .97 for the Grade 5 assessment and .96 for the Grade 6 assessment. For participants with more than one reciprocated friend, an average score for all friends was computed. The 303 participants (30.4%) who had no reciprocated friends at both assessments did not have to be excluded altogether from the main analyses, as the full information maximum likelihood method was used to manage missing data. More information on this procedure is provided in the *Attrition* subsection under the Results section. A correlation of .22 ( $p < .001$ ) was found for reciprocated friends' scores of aggression-disruptiveness in Grades 5 and 6 and was based on participants with complete



data on their reciprocated friends' characteristics at both times of measurement (38.4% of the total sample).

*Academic achievement and school commitment during secondary school.* Teachers reported on students' academic achievement in French (first language) and mathematics from 1992 through 1995 (when the students were ages 14 through 17), which correspond to Grades 8 through 11 for students in age-appropriate classrooms.<sup>2</sup> An average score of academic achievement in French and mathematics across the 4 years was computed to represent academic achievement during adolescence. Final scores ranged on a continuous scale from 0 (*academic failure*) to 4 (*excellent academic performance*). Correlations between scores of academic achievement for all pairs of assessments over this 4-year period ranged from .29 to .58 ( $p < .001$ ). The proportion of participants with complete data on each pair of assessments ranged from 60.3% to 76.9% of the total sample.

School commitment was measured through self-reports with a seven-item scale including such statements as "How much does having good grades matter to you?" and "How long do you intend to stay in school?" Once again, an average score, based on all available data gathered from 1992 through 1995 (when the students were ages 14 through 17), was used. Each item was rated on a scale ranging from 1 to 4 (total scores ranging from 7 to 28), with higher scores reflecting higher school commitment. Cronbach's alphas for this scale ranged from .67 to .71, depending on the year of assessment. Correlations between scores of school commitment for all pairs of assessments over this 4-year period ranged from .47 to .69 ( $p < .001$ ). The proportion of participants with complete data on each pair of years ranged from 63.6% to 77.4% of the total sample.

*Graduation from secondary school.* Data on graduation status by early adulthood (when the students were age 23; December 2001) were available from the Education Department's official records. Although the normal age of graduates in the province of Quebec is 17 years, temporary dropout and grade retention delayed the time of graduation for several students. In the current sample, only 32.2% of the participants were placed in regular, age-appropriate classrooms by the time they reached their final year of secondary schooling, and 11.2% of the participants were 1 year behind. The remaining 56.6% of the participants were 2 or more years behind grade level, had been placed in special classrooms for students with learning or behavioral difficulties, or had dropped out of secondary school. Still, by 23 years of age, 26.3% of this final group had graduated (89.9% of students who were no more than 1 year behind had graduated by age 23). According to Bowlby and McMullen (2005), the probability of completing secondary schooling has dramatically decreased by 20 to 24 years old. Age 23 is thus an appropriate time for measuring graduation status, as this status is likely to be definitive for most participants.

### Procedure

Participants were recruited around the end of their kindergarten year (in the spring of 1984). With the collaboration of 53 public schools located in low-SES areas of Montreal and its immediate surroundings, we sent questionnaires to the parents of all boys attending kindergarten at those schools. A consent form explaining the study was sent in the same package and had to be signed and returned by mail along with the questionnaire. This procedure was

repeated every year, and parents were informed that their son would fill out questionnaires at school when that was the case. For each participant, one teacher filled out a questionnaire assessing the student's behavior and academic achievement at every year.

### Results

Structural equation modeling (SEM) was used to assess the fit of the model presented in Figure 1. Attrition, intraclass correlations (to assess between-schools effects), and bivariate correlations among the main measures are presented first.

#### Preliminary Analyses

*Attrition.* The use of SEM allowed for the inclusion of all participants having complete data on exogenous variables, even if they had missing data on other variables. In the current model, SES was the only exogenous variable. It was therefore important to minimize the number of participants with missing data on this particular variable. Data on SES at the first wave of data collection (1984) were available for 973 participants, but given that the correlation between SES in 1984 and the same variable measured in subsequent years was high (ranging from .62 to .67), available data from the nearest following year (1988, 1989, or 1990) were used for participants with missing values in 1984. Using this strategy, only 40 participants were lost due to missing data. Thus, 997 participants out of the 1,037 initially recruited could be included in the primary analysis using SEM, preserving 96.1% of the initial sample.

Excluded participants had significantly lower academic achievement in elementary school,  $t(967) = -3.93, p < .001$ , lower levels of prosocial behaviors,  $t(1035) = -2.82, p < .01$ , lower school commitment in secondary school,  $t(901) = -2.09, p < .05$ , and lower academic achievement in secondary school,  $t(896) = -3.19, p < .01$ , than participants who were included in the primary analysis. A chi-square test also revealed that excluded participants tended to be less likely to graduate from secondary school, although this difference was only marginally significant,  $\chi^2(1, N = 1,018) = 3.67, p = .06$ . However, no significant differences emerged on their levels of aggressive-disruptive behaviors in childhood, their friends' levels of aggressive-disruptive behaviors, and their levels of peer acceptance.

Missing data on endogenous variables were handled using the full information maximum likelihood method. Therefore, the covariance matrix was constructed from all available information for each participant. Complete data on each relevant pair of variables ranged from 67.8% to 100.0% of the final sample ( $N = 997$ ), with an average of 86.5%.

*Intraclass correlations.* Because participants were recruited from different schools, intraclass correlations were computed in order to verify whether some of the variance in the main variables of the model (i.e., participants' behaviors in childhood, peer experiences in preadolescence, academic achievement and school commitment in adolescence, and failure to graduate) could be

<sup>2</sup> In contrast with the school systems in other Canadian provinces and in the United States, secondary schooling in the province of Quebec lasts only 5 years, thus ending in Grade 11. As a result, students normally enter secondary school at 12 years old and graduate at 17 years old.

attributed to a higher order "school effect." These analyses revealed that a significant proportion of variance in these variables was attributable to school. Effect sizes were small to moderate (Hox, 2002), ranging from 4% (for aggressive-disruptive behaviors in childhood) to 11% (for failure to graduate). Although the hierarchical structure of the data may suggest that a multilevel analytic framework is desirable, the distribution of the participants into 197 different institutions after the transition to secondary school entails some restrictions with regard to the application of a multilevel analytical procedure. In fact, the small number of participants per school—50% of the schools included only 1 participant, and only 25% of schools included 5 or more participants—made implementing a multilevel framework inadvisable (Newsom & Nishishiba, 2002).

**Bivariate correlations.** Correlations, which are presented in Table 2, were all in the expected direction. First, SES was positively related to academic achievement, prosocial behavior, peer acceptance, and school commitment but negatively related to participants' aggressive-disruptive behaviors and failure to graduate. However, SES was unrelated to friends' level of aggressive-disruptive behaviors.

Still in line with the hypotheses, academic results in childhood were negatively correlated with concurrent aggressive-disruptive behaviors but positively associated with prosocial behaviors. In addition, boys with higher levels of academic achievement in childhood tended to associate with less aggressive-disruptive friends in preadolescence. They also tended to have higher levels of peer acceptance in preadolescence, higher levels of school commitment and academic achievement during adolescence, and lower risks of the failure to graduate.

Aggressive-disruptive behavior in childhood was negatively correlated with prosocial behaviors. Boys who were more aggressive-disruptive in childhood tended to associate with aggressive-disruptive friends and to have low levels of peer acceptance in preadolescence. Early aggression-disruptiveness was also related to low levels of adolescent school commitment and academic achievement and to the failure to graduate from secondary school. Prosocial boys tended to experience greater peer ac-

ceptance and to have friends with lower levels of aggressive-disruptive behaviors in preadolescence. Prosociality was also related to higher levels of school commitment and academic achievement during adolescence as well as to a lower risk of the failure to graduate.

No correlation emerged between peer acceptance and the concurrent level of friends' aggressive-disruptive behaviors. Still, having more aggressive-disruptive friends in preadolescence predicted lower levels of adolescent school commitment and academic achievement as well as a greater risk of the failure to graduate. In contrast, greater peer acceptance in preadolescence predicted greater adolescent school commitment and academic achievement as well as a lower risk of the failure to graduate.

Finally, greater school commitment was positively related to academic achievement in adolescence. Both commitment and academic achievement were inversely related to the failure to graduate.

### Model Testing

All analyses were conducted with *Mplus* Version 3.01 (Muthén & Muthén, 2004) using weighted least squares estimation. We first examined the fit of a baseline model in which all variables were completely unrelated. The fit of this model was very poor,  $\chi^2(20, N = 997) = 1,050.45, p < .001$ . (In SEM analyses, a good model fit usually yields a nonsignificant chi-square statistic). We then tested our hypothesized model in order to assess any improvements in fit.

The hypothesized model ( $M_1$ )—a saturated model—was tested first. Some of the hypothesized paths were not significant and were therefore removed in order to build a more parsimonious model. The final model,  $M_2$ , is illustrated in Figure 2, with standardized coefficients for all significant paths. Arrows in bold indicate significant indirect pathways. Fit indices suggest that this model fits the data very well,  $\chi^2(10, N = 997) = 8.57, p = .57$ , comparative fit index = 1.00, Tucker-Lewis index = 1.00 (comparative fit index and Tucker-Lewis index values above .95 are indicative of a good fit). The nonsignificant chi-square value of the final model

Table 2  
*Bivariate Correlations for All Variables*

Measure	1	2	3	4	5	6	7	8	9
Control variables									
1. SES	—								
2. Academic achievement	.30***	—							
Period 1									
3. Aggression-disruptiveness	-.19***	-.35***	—						
4. Prosociality	.10**	.24***	-.25***	—					
Period 2									
5. Friends' aggression-disruptiveness	.00	-.08*	.09*	-.09*	—				
6. Peer acceptance	.15***	.21***	-.25***	.22***	-.01	—			
Period 3									
7. School commitment	.23***	.24***	-.21***	.19***	-.16***	.12***	—		
8. Academic achievement	.29***	.46***	-.26***	.13***	-.17***	.14***	.38***	—	
Period 4									
9. Failure to graduate	-.35***	-.54***	.36***	-.18***	.12**	-.18***	-.41***	-.50***	—

Note. Period 1 corresponds to childhood (ages 6 or 7 and 10 years), Period 2 corresponds to preadolescence (ages 11 and 12 years), Period 3 corresponds to adolescence (ages 14 through 17 years), and Period 4 corresponds to early adulthood (age 23 years). SES = socioeconomic status.  
\* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .



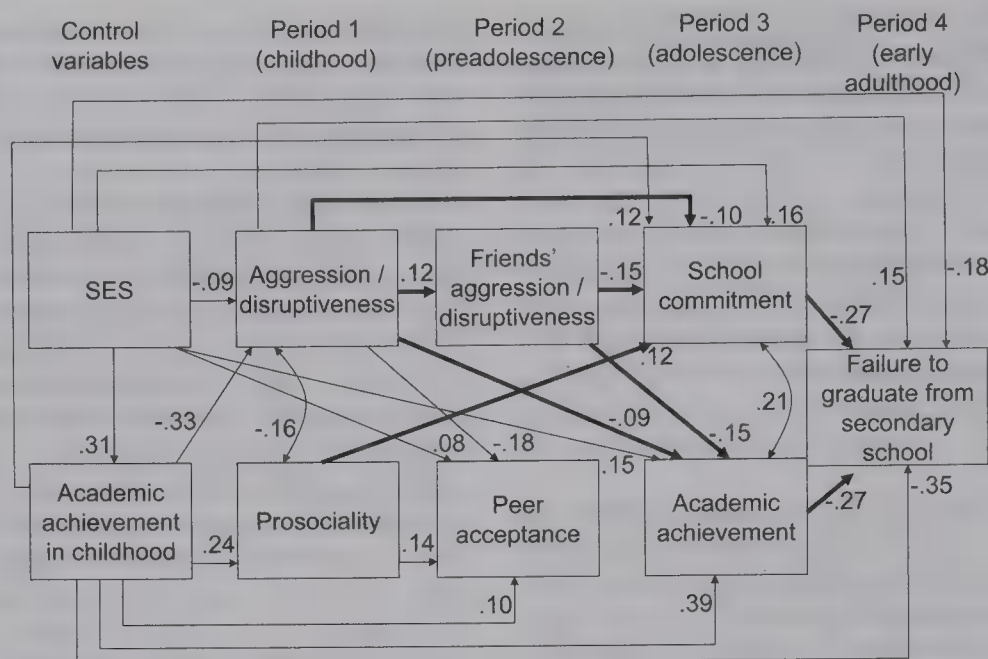


Figure 2. Final model. Numbers indicate significant standardized covariance estimates. Arrows in bold indicate indirect paths. SES = socioeconomic status.

( $M_2$ ) suggests that the original, saturated model ( $M_1$ ) does not fit the data significantly better than does the more parsimonious model ( $M_2$ ).

**Direct paths.** In line with the hypotheses, most of the direct paths included in the initial model were statistically significant.

First, SES was a significant predictor of childhood academic achievement. In line with our hypotheses, higher SES predicted lower levels of childhood aggressive-disruptive behavior, higher preadolescent peer acceptance, and higher adolescent academic achievement and school commitment. Higher SES was associated with lower odds of the failure to graduate.

Higher levels of the other control variable, childhood academic performance, predicted lower levels of aggressive-disruptive behavior and higher prosociality in childhood. It also predicted higher preadolescent peer acceptance and higher adolescent academic achievement and school commitment. Finally, it was negatively associated with the failure to graduate.

Childhood aggression-disruptiveness was negatively associated with concurrent levels of prosociality. It was positively associated with friends' preadolescent aggression-disruptiveness and negatively associated with preadolescent peer acceptance. Higher levels of childhood aggression-disruptiveness were associated with lower adolescent academic achievement and school commitment, and it was a significant predictor of the failure to graduate. In contrast, childhood prosociality predicted higher preadolescent peer acceptance and adolescent school commitment.

As revealed by the preliminary bivariate analyses, preadolescent peer acceptance and friends' aggression-disruptiveness were not significantly intercorrelated. Friends' preadolescent aggression-disruptiveness predicted lower levels of adolescent academic achievement and school commitment. However, in contrast to the bivariate analyses, preadolescent peer acceptance was not related to the hypothesized adolescent mediators (i.e., school commitment and academic achievement).

Finally, both adolescent mediators, school commitment and academic achievement, were significantly and positively intercor-

related. They were associated with lower odds of the failure to graduate.

**Indirect paths.** The principal objective of this study was to determine whether peer experiences played an active (i.e., mediating) role in the pathways leading to the failure to graduate from secondary school. We tested a number of indirect paths in which peer experiences were hypothesized (a) to relate to the failure to graduate via school-related variables in adolescence and (b) to mediate the association between childhood behaviors and the adolescent precursors of the failure to graduate.

Eight indirect paths were modeled in order to reflect all possible indirect pathways from childhood aggressive-disruptive and prosocial behaviors to the failure to graduate from secondary school in adulthood via friends' aggressive-disruptive behaviors and peer acceptance in preadolescence and subsequent adolescent school commitment and academic achievement. In addition, we tested four indirect paths from childhood aggressive-disruptive and prosocial behaviors to the failure to graduate that included school commitment and academic achievement as mediators but excluded the peer experiences. This procedure was used to minimize the risk of overestimating the variance attributed to the indirect paths that included the peer experiences.

The final model revealed support for two of the eight possible indirect paths involving peer experiences. First, as illustrated in Figure 2 by the bold arrows, the path starting from participants' aggressive-disruptive behaviors and running through friends' aggressive-disruptive behaviors as well as school commitment was significant, although the standardized coefficient for this pathway was small (.01). The other significant indirect path started from participants' aggressive-disruptive behavior, which led to friends' aggressive-disruptive behaviors, adolescent academic achievement, and finally, the failure to graduate (standardized path coefficient = .01). Peer acceptance was not a mediator in any pathway linking early behavior to the failure to graduate.

It is noteworthy that in addition to the role played by aggressive-disruptive behavior in childhood in the above-mentioned indirect

pathways involving friends' aggressive-disruptive behavior, participants' aggression-disruptiveness was also related to the failure to graduate through other pathways. For example, childhood aggression-disruptiveness was directly related to adolescent school commitment which, in turn, predicted the failure to graduate (standardized path coefficient = .03). Childhood aggression-disruptiveness was also related to adolescent academic achievement which, in turn, predicted the failure to graduate (standardized path coefficient = .02).

One additional indirect path—from prosocial behavior in childhood to school commitment to the failure to graduate—explained a significant proportion of variance in this outcome (standardized path coefficient = -.03).

## Discussion

The current study was meant to test a comprehensive model of the failure to graduate from secondary school in which two types of peer experiences—peer acceptance and friends' aggressive-disruptive behavior—were hypothesized (a) to be independent and significant predictors of the failure to graduate and (b) to play a mediating role in the path from early behavior to the failure to graduate. Specifically, as suggested by the social interactional perspective, peer experiences were embedded into a chain of events occurring over several developmental periods and were hypothesized to be predicted by behavior patterns in childhood. Also, according to the participation-identification perspective, two proximal predictors of the failure to graduate were hypothesized to mediate the link between peer experiences during preadolescence and secondary school graduation, including school commitment and academic achievement during adolescence.

### *SES, Early Academic Performance, and Childhood Behavior Patterns*

Although peer experiences were the central variables in our model, several of the links among variables that were included in the model for control purposes were significant and warrant attention. First, as expected, SES had a significant direct and negative relationship with the failure to graduate. As explained earlier, parents' behaviors and expectations, together with the lack of family resources, are potential mediators of this link. However, these variables were not included in the study, which might account for the residual link between SES and the failure to graduate.

Second, poor academic achievement in the early elementary school years was found to predict the failure to graduate over and above academic achievement in secondary school. Although proximal measures of a given variable (in this case, academic achievement in secondary school) are usually expected to be stronger predictors of the outcome, the current study shows that this is not always true. Our finding, however, makes sense from a developmental psychopathology perspective. Early schooling experiences initially direct students along an academic trajectory and may lead to a range of events that keep students on that particular path. In the current study, poor academic achievement during the first years of schooling not only was the strongest direct predictor of the failure to graduate but also contributed to other variables in the model (i.e., participants' aggressive-disruptive and prosocial behaviors in childhood and academic achievement and commitment

in secondary school). Early academic difficulties might be stronger predictors of the failure to graduate than later difficulties because they affect young children's school-related self-esteem and self-efficacy as well as their relationships with school staff. In contrast, academic difficulties that emerge in secondary school may simply precipitate school dropout in some students without affecting students' deeper beliefs in their academic abilities or capacity to graduate if they ever wish to complete the secondary school program.

Third, a significant, direct link was found between aggressive-disruptive behaviors in childhood and the failure to graduate by early adulthood. This link, which was also found in several empirical studies, cannot be entirely explained by ensuing peer difficulties. Other mediators should therefore be explored in future studies. For example, being more aggressive and disruptive than other students may be the first step in a developmental pathway toward the failure to graduate because such behaviors affect students' relationships with teachers. These students may not receive the same amount of help, support, and encouragement in their schoolwork (Brendgen, Wanner, Vitaro, Bukowski, & Tremblay, 2007).

In contrast to participants' levels of aggression-disruptiveness, no residual link emerged between prosociality in childhood and the failure to graduate from secondary school. The contribution of low prosociality to the failure to graduate was therefore entirely mediated by school commitment in adolescence. This suggests that prosocial behaviors in school may be an early indicator of a student's stable tendency to conform to social norms and expectations. In addition, this finding could indicate that prosocial children are more likely to have positive experiences in school (e.g., being praised by the teacher) that help them identify with the institution.

### *The Role of Aggressive-Disruptive Friends*

In partial support of our hypotheses, two of the indirect pathways involving the association with aggressive-disruptive friends during preadolescence were significant. Indeed, both pathways starting from aggression-disruptiveness in childhood and running through friends' aggressive-disruptive behavior during preadolescence were significant. The first pathway contributed to school graduation through the link between friends' behavior and school commitment, whereas the second pathway reached school graduation through the link between friends' behavior and academic achievement. These findings support the validity of our integrative theoretical model, in which variables from the participation-identification perspective (i.e., school commitment and academic achievement) effectively complement the social interactional framework.

In contrast, the indirect pathways involving prosocial skills in childhood, peer experiences, and graduation status by early adulthood were nonsignificant. This is surprising, given that a lack of prosocial skills has been hypothesized to contribute to negative peer experiences to the same degree as aggressive-disruptive behaviors (Patterson et al., 1989). Still, we found an indirect pathway starting from low prosociality in childhood and predicting the failure to graduate through its link with low school commitment in adolescence. It seems possible that other types of peer-related problems that were not measured in the current study, such as



friendlessness or an association with friends who are deviant but not aggressive-disruptive (e.g., friends who are truant or who cheat on exams but who are not disruptive in the classroom and who do not get involved into fights), are involved in this indirect pathway.

In line with this suggestion, future studies should extend their measures of friends' characteristics to include not only negative, but also positive traits, as previous research has shown that friends can play a positive role in the context of students' school adjustment (Berndt, Hawkins, & Jiao, 1999; Berndt & Keefe, 1995; Chen, 2005; Mounts & Steinberg, 1995; Wentzel, Barry, & Caldwell, 2004). For example, affiliating with friends who have developed positive attitudes toward school and authority figures, who intend to undertake postsecondary education, and who display good study habits is likely to predict later secondary school completion. Affiliation with such friends could also protect vulnerable students against school disengagement and the failure to graduate.

The current study's results suggest that parents of preadolescent boys who are at risk for the failure to graduate can promote their sons' successful completion of the secondary school program by monitoring the boys' social activities during preadolescence. Although parents often allow their children to spend more time with their friends in unsupervised settings as they grow older, mothers and fathers should at least get to know their children's friends and encourage associations with those who do not display excessive levels of aggression and disruptiveness. This is especially important in the case of boys who have a personal history of aggressive-disruptive behavior or who displayed low levels of academic achievement in elementary school.

Furthermore, because many friendships arise in the school setting, teachers are in a very good position to identify potentially risky associations between aggressive-disruptive youngsters. Teachers should work together with parents to encourage at-risk students to become friends with well-adapted students who, in addition to being positive role models, may also be able to actively discourage deviant behaviors and negative attitudes toward school. School administrators can contribute to this end by providing a variety of extracurricular activities through which at-risk students can meet and develop friendships with normative peers in adult-supervised settings.

### *The Role of Peer Acceptance*

In contrast to the positive results obtained for the other type of peer experience, the role of peer acceptance as a fundamental element in the developmental pathway leading to school graduation was not supported by the current study. Still, peer acceptance should not be dismissed as a potential contributor to school graduation.

First, the timing of measurement might explain the failure of peer acceptance to play a role in the pathway leading to school graduation. In fact, although peer acceptance should, according to some authors (Buhrmester & Furman, 1986), contribute to psychological adjustment even after childhood, others suggested that peer acceptance is most important during the first years of schooling. In contrast, dyadic peer experiences (such as the association with aggressive-disruptive friends) are most important in preadolescence or adolescence, according to this second group of authors (Patterson et al., 1992; Sullivan, 1953).

Second, as suggested by the social interactional perspective, early peer rejection may pave the way for later involvement with aggressive-disruptive friends. To test for this possibility, future studies of the developmental processes leading to secondary school graduation should adopt a strategy that would pit two models against one another. In one model, peer acceptance and association with aggressive-disruptive friends would act on future adjustment in a parallel manner. In the alternative sequential model, peer rejection would predict subsequent associations with deviant friends, which would in turn predict school adjustment problems in early adulthood. Comparing these two models would require the assessment of both types of peer experiences throughout childhood and adolescence. As we did not measure peer experiences before and after preadolescence, such a model could not be tested in the current study.

A third explanation for the nonsignificant results obtained with peer acceptance is that the operationalization of this variable might have failed to represent the construct it was meant to measure. In fact, in the theories proposed by Sullivan (1953) and by Patterson and his colleagues (1992), peer acceptance corresponds to being well-liked by one's peers. Students' positive feelings toward one another are best measured with a procedure of like-most and like-least peer nominations (e.g., Coie, Dodge, & Coppotelli, 1982; Newcomb & Bukowski, 1983). In the current study, however, peer acceptance was evaluated through peer nominations of students having many friends and of those having very few friends. According to Parkhurst and Hopmeyer's (1998) criteria, this procedure may be closer to an assessment of "peer-perceived popularity," rather than "sociometric popularity." In other words, our measure probably reflected the extent to which classmates considered our participants as popular and high in social status, rather than their true feelings of liking toward our participants. Indeed, Parkhurst and Hopmeyer found that peer-perceived popularity is a correlate of social dominance, and although some students who get high ratings on this measure are perceived as kind and trustworthy by their peers, many others are perceived as aggressive and "stuck-up." Other researchers who measured peer-perceived popularity by asking participants to rate their classmates on a 5-point scale ranging from *not at all popular* to *very popular* found that increases in this variable over four academic semesters were related to decreases in academic achievement over the same period of time, but only in highly aggressive adolescents (Schwartz, Gorman, Nakamoto, & McKay, 2006). In contrast, Ollendick, Weist, Borden, and Greene (1992) reported that a rejected status, measured through sociometric nominations, was associated with a higher risk of dropping out of school by Grade 9.

It is noteworthy that the significant results obtained with friends' aggressive-disruptive behaviors and the nonsignificant results obtained with peer acceptance may suggest the existence of a third variable effect—also known as a spurious or incidental effect (Ladd & Troop-Gordon, 2003; Woodward & Fergusson, 2000). In fact, one might put forward the hypothesis that any significant relationship between peer acceptance and later school adjustment found in previous studies was a mere reflection of the true relationships existing between two correlates of peer acceptance, namely, friends' aggressive-disruptive behaviors and school adjustment. However, a significant relationship between friends' aggression-disruptiveness and peer acceptance would have been necessary to support such an explanation, and because the corre-



lation between these variables was null (even in the bivariate analyses), there is little support for the third variable hypothesis.

### Limitations

First, some of the variables that may relate to graduation were not included in the model, and this limits our understanding of the overall process leading to the failure to graduate. This study focused on peer acceptance and friends' aggression-disruptiveness as key potential predictors of the failure to graduate. Therefore, other peer-related variables were not addressed here, including victimization by peers, crowd affiliation (i.e., the identification with a reputation-based group of peers like the "popular students," the "druggies," or the "brains"), clique membership (i.e., the association with a group of close friends sharing similar traits, values, and activities), clique-related status (e.g., central or peripheral), friendlessness, and friendship quality. These variables represent a range of peer experiences that could be related to concurrent and future academic outcomes.

In addition to peer-related variables, family-related variables may also enrich our understanding of the factors that influence graduation outcomes. This study incorporated only family SES. Yet, as suggested by the social interactional perspective, other family variables may influence graduation status, including parental supervision—especially at the time when children enter adolescence and become more independent from parents (Patterson et al., 1992). Other researchers have called attention to the quality of early parenting and home environment (Jimerson, Egeland, Sroufe, & Carlson, 2000), as well as to parents' involvement in their children's schooling, expectations for their children's educational attainment, encouragement of the children's achievement, relationships with school staff, and participation in school activities as potential predictors of later academic outcomes (Astone & McLanahan, 1991; Kohl, Lengua, & McMahon, 2000).

The social interactional perspective also suggests that the establishment of a positive relationship with the teacher may contribute to students' school adjustment (Brendgen et al., 2007; Pianta & Stuhlman, 2004). Further, recent reviews point to structural and human characteristics of the school that may influence graduation rates (e.g., school and class size; the availability of alternative academic curricula for potential dropouts; the valuing of work and learning; the adequacy of educative materials; positive relationships among school staff members, parents, and students; and the availability of extracurricular activities; Baker et al., 2001; Rutter & Maughan, 2002). These hypothesized relations are also consistent with the participation-identification view, inasmuch as a positive school climate is thought to encourage active participation in school activities and should make it easier for students to identify with the institution's goals and values (Finn, 1989).

Second, the methodological limitations of this study also deserve some consideration. This study focused on a high-risk population: boys from low-SES, urban neighborhoods. Although it is important to understand the pathway toward the failure to graduate taken by these vulnerable students, this limits the generalizability of our findings. More research will be necessary in order to verify whether the current findings are also true for girls, for students of other ethnic groups, for youngsters living in medium-to-high-SES neighborhoods, and for those living in rural or semirural areas.

As previously mentioned, the nested structure of our sample (i.e., clusters of students recruited from several schools) raises questions about the comparability of the school experience across participants attending different institutions. The significant intra-class correlations reported in the Results section support the idea that school characteristics may, to some extent, influence students' academic achievement, school behaviors, attitudes toward schooling, and ultimately, chances of graduating. Future studies involving several schools should therefore include a sufficient number of participants per school to allow for a more thorough investigation of the schools' contributions to students' developmental trajectories through multilevel modeling.

In addition, the current study tested a hybrid model based on the combination of sequential and parallel mediation pathways. Such longitudinal studies aid in the elucidation of long-term developmental processes, including those leading to school graduation. Currently, however, bidirectional models (also known as cross-lagged or transactional models) are becoming increasingly popular as researchers recognize that the chain of events leading to a particular developmental outcome is likely to be much more complex than a simple "A causes B which causes C" relationship. Although short-term longitudinal models involving reciprocal effects among variables have been tested (e.g., Welsh, Parke, Widaman, & O'Neil, 2001), long-term longitudinal models of this kind are still scarce. The database used in the current study did not allow us to test for reciprocal effects because many of the relevant variables were not measured at all developmental periods. Nevertheless, future studies based on more recent data sets can answer important questions regarding the short-term, bidirectional mediation processes that comprise the larger pathway toward secondary school graduation.

The issue of missing data is inherent to longitudinal study designs. In the current study, we addressed the issue by using all available data over a specific developmental period for most of the variables. More specifically, average scores based on two assessments were used for the variables measured in Periods 1 and 2, and average scores based on four assessments were used for the variables measured in Period 3. Still, this strategy has its own drawbacks. Although correlations between the scores obtained at different assessments were highly significant, the size of the correlation was quite small for the prosociality measure. Because children's prosociality scores had a general tendency to drop between kindergarten and Grade 4, it seems plausible that the measure of prosociality itself was not equally suitable for measuring this variable at both ages—possibly because the behavioral manifestations of prosociality differ in younger and older children. Low correlations were also found for the measure of aggression-disruptiveness in participants' reciprocated friends over a 1-year interval. Because this measure was based on peer nominations and standardized within classrooms, the low correlation between the two assessment points probably reflects construct instability—that is, the formation of new friendships with students having different levels of aggression-disruptiveness. Aggregating weakly correlated measurements of the same construct may have negatively affected the power of the SEM analysis.

One last limitation of this study concerns the small effect sizes of the significant indirect pathways. The standardized path coefficients of these indirect pathways were all below .10, which is the benchmark for a small effect size (Cohen, 1988). These small



effect sizes were not unexpected given that many mediators were included in the model. Still, every student had his own history of social and academic experiences. Each significant pathway is thus likely to be a worthwhile explanation of the course taken by a small but significant number of individuals during their academic careers, so the indirect pathways presented here should not be dismissed simply on the basis of their small effect sizes. The limited reliability of the scores obtained for some variables (i.e., participants' prosociality and friends' aggression-disruptiveness), as previously discussed, may also have contributed to small effect sizes.

## Conclusion

The current study sheds new light on peers' contribution to the developmental pathway leading to the failure to graduate. In particular, this research highlighted the negative relation between reciprocated friendships with aggressive-disruptive peers and the likelihood of secondary school graduation, which can be attributed to the mediating roles of school commitment and academic achievement.

Although the many questions raised in the Discussion show that much research still needs to be done in order to achieve a full understanding of the role of peer experiences on the pathway leading to secondary school graduation, the current study can inform the development of experimental prevention programs aimed at promoting secondary school completion. The clear association of early academic failure and behavior problems with the failure to graduate from secondary school suggests that interventions should target the development of social skills and school readiness before school entry (e.g., Schweinhart, Barnes, & Weikart, 1993). Furthermore, experimental prevention programs represent good settings to develop strategies that will help at-risk preadolescent students form and maintain friendships with normative peers, perhaps by teaching at-risk preadolescents more appropriate social skills (Asher, Parker, & Walker, 1996) and by involving at-risk students in extracurricular activities that are supervised by well-trained adults and in which well-adjusted youngsters are also taking part (Silver & Eddy, 2006).

Finally, because not all at-risk students have access to structured prevention programs, parents and teachers should be encouraged to monitor closely the peer relationships of at-risk boys and to offer these students opportunities to make friends with normative, prosocial peers. This might be just what it takes for some students to get on the pathway to secondary school graduation, thereby reducing their risk of future psychosocial maladjustment.

## References

- Asher, S. R., Parker, J. G., & Walker, D. L. (1996). Distinguishing friendship from acceptance: Implications for intervention and assessment. In W. M. Bukowski, A. F. Newcomb, & W. W. Hartup (Eds.), *The company they keep: Friendship in childhood and adolescence* (pp. 366–403). Cambridge, United Kingdom: Cambridge University Press.
- Astone, N. M., & McLanahan, S. S. (1991). Family structure, parental practices and high school completion. *American Sociological Review*, 56(3), 309–320.
- Bagwell, C. L., & Coie, J. D. (2004). The best friendships of aggressive boys: Relationship quality, conflict management, and rule-breaking behavior. *Journal of Experimental Child Psychology*, 88(1), 5–24.
- Baker, J. A., Derrer, R. D., Davis, S. M., Dinklage-Travis, H. E., Linder, D. S., & Nicholson, M. D. (2001). The flip side of the coin: Understanding the school's contribution to dropout and completion. *School Psychology Quarterly*, 16(4), 406–426.
- Battin-Pearson, S., Newcomb, M. D., Abbott, R. D., Hill, K. G., Catalano, R. F., & Hawkins, J. D. (2000). Predictors of early high school dropout: A test of five theories. *Journal of Educational Psychology*, 92(3), 568–582.
- Berndt, T. J., Hawkins, J. A., & Jiao, Z. (1999). Influences of friends and friendships on adjustment to junior high school. *Merrill-Palmer Quarterly*, 45(1), 13–41.
- Berndt, T. J., & Keefe, K. (1995). Friends' influence on adolescents' adjustment to school. *Child Development*, 66(5), 1312–1329.
- Blisshen, B. R., Carroll, W. K., & Moore, C. (1987). The 1981 socioeconomic index for occupations in Canada. *Canadian Review of Sociology and Anthropology*, 24, 465–488.
- Bowlby, G., & McMullen, K. (2005). *Provincial dropout rates—Trends and consequences* (No. 81–004-XIE). Ottawa, Ontario, Canada: Statistics Canada.
- Brendgen, M., Vitaro, F., & Bukowski, W. M. (1998). Affiliation with delinquent friends: Contributions of parents, self-esteem, delinquent behavior, and rejection by peers. *Journal of Early Adolescence*, 18(3), 244–265.
- Brendgen, M., Wanner, B., Vitaro, F., Bukowski, W. M., & Tremblay, R. E. (2007). Verbal abuse by the teacher during childhood and academic, behavioral, and emotional adjustment in young adulthood. *Journal of Educational Psychology*, 99(1), 26–38.
- Buhrmester, D., & Furman, W. (1986). The changing functions of friends in childhood: A neo-Sullivanian perspective. In V. J. Derlega & B. A. Winstead (Eds.), *Friendship and social interaction* (pp. 41–62). New York: Springer-Verlag.
- Burk, W. J., & Laursen, B. (2005). Adolescent perceptions of friendship and their associations with individual adjustment. *International Journal of Behavioral Development*, 29(2), 156–164.
- Chen, J. J. L. (2005). Relation of academic support from parents, teachers, and peers to Hong Kong adolescents' academic achievement: The mediating role of academic engagement. *Genetic, Social, and General Psychology Monographs*, 131(2), 77–127.
- Christenson, S. L., Sinclair, M. F., Lehr, C. A., & Godber, Y. (2001). Promoting successful school completion: Critical conceptual and methodological guidelines. *School Psychology Quarterly*, 16(4), 468–484.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Coie, J. D., Dodge, K. A., & Coppotelli, H. (1982). Dimensions and types of social status: A cross-age perspective. *Developmental Psychology*, 18(4), 557–570.
- Coie, J. D., Dodge, K. A., & Kupersmidt, J. B. (1990). Peer group behavior and social status. In S. R. Asher & J. D. Coie (Eds.), *Peer rejection in childhood* (pp. 17–59). New York: Cambridge University Press.
- Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment. *Journal of Family Psychology*, 19(2), 294–304.
- Dishion, T. J., Andrews, D. W., & Crosby, L. (1995). Antisocial boys and their friends in early adolescence: Relationship characteristics, quality, and interactional process. *Child Development*, 66(1), 139–151.
- Dishion, T. J., McCord, J., & Poulin, F. (1999). When interventions harm: Peer groups and problem behavior. *American Psychologist*, 54(9), 755–764.
- Dishion, T. J., Patterson, G. R., Stoolmiller, M., & Skinner, M. L. (1991). Family, school, and behavioral antecedents to early adolescent involvement with antisocial peers. *Developmental Psychology*, 27(1), 172–180.
- Education Department. (2001). *Indicateurs de l'éducation, Édition 2002*

- [Education indicators, 2002 edition] (No. 2002-01-01209). Quebec City, Quebec, Canada: Gouvernement du Québec.
- Ekstrom, R. B., Goertz, M. E., Pollack, J. M., & Rock, D. A. (1986). Who drops out of high school and why? Findings from a national study. *Teachers College Record*, 87(3), 356-373.
- Entwisle, D. R., Alexander, K. L., & Olson, L. S. (2004). Temporary as compared to permanent high school dropout. *Social Forces*, 82(3), 1181-1205.
- Finn, J. D. (1989). Withdrawing from school. *Review of Educational Research*, 59(2), 117-142.
- Furman, W., & Robbins, P. (1985). What's the point? Issues in the selection of treatment objectives. In B. H. Schneider, K. H. Rubin, & J. E. Ledingham (Eds.), *Children's peer relations: Issues in assessment and intervention* (pp. 41-56). New York: Springer-Verlag.
- Gifford-Smith, M. E., & Brownell, C. A. (2003). Childhood peer relationships: Social acceptance, friendships, and peer networks. *Journal of School Psychology*, 41(4), 235-284.
- Hartup, W. W. (1996). The company they keep: Friendships and their developmental significance. *Child Development*, 67(1), 1-13.
- Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.
- Hymel, S., Comfort, C., Schonert-Reichl, K., & McDougall, P. (1996). Academic failure and school dropout: The influence of peers. In J. Juvonen & K. R. Wentzel (Eds.), *Social motivation: Understanding children's school adjustment* (pp. 313-345). New York: Cambridge University Press.
- Jimerson, S. R., Egeland, B., Sroufe, L., & Carlson, B. (2000). A prospective longitudinal study of high school dropouts: Examining multiple predictors across development. *Journal of School Psychology*, 38(6), 525-549.
- Kandel, D. B. (1978). Homophily, selection, and socialization in adolescent friendships. *American Journal of Sociology*, 84(2), 427-436.
- Kaplan, D. S., Damphousse, K. R., & Kaplan, H. B. (1994). Mental health implications of not graduating from high school. *Journal of Experimental Education*, 62(2), 105-123.
- Kerckhoff, A. C., & Bell, L. (1998). Early adult outcomes of students at "risk." *Social Psychology of Education*, 2(1), 81-102.
- Kohl, G. O., Lengua, L. J., & McMahon, R. J. (2000). Parent involvement in school: Conceptualizing multiple dimensions and their relations with family and demographic risk factors. *Journal of School Psychology*, 38(6), 501-523.
- Ladd, G. W., Kochenderfer, B. J., & Coleman, C. C. (1996). Friendship quality as a predictor of young children's early school adjustment. *Child Development*, 67(3), 1103-1118.
- Ladd, G. W., & Troop-Gordon, W. (2003). The role of chronic peer difficulties in the development of children's psychological adjustment problems. *Child Development*, 74(5), 1344-1367.
- Laird, R. D., Jordan, K. Y., Dodge, K. A., Pettit, G. S., & Bates, J. E. (2001). Peer rejection in childhood, involvement with antisocial peers in early adolescence, and the development of externalizing behavior problems. *Development and Psychopathology*, 13(2), 337-354.
- Lansford, J. E., Criss, M. M., Pettit, G. S., Dodge, K. A., & Bates, J. E. (2003). Friendship quality, peer group affiliation, and peer antisocial behavior as moderators of the link between negative parenting and adolescent externalizing behavior. *Journal of Research on Adolescence*, 13(2), 161-184.
- McCaul, E. J., Donaldson, G. A., Jr., Coladarc, T., & Davis, W. E. (1992). Consequences of dropping out of school: Findings from high school and beyond. *Journal of Educational Research*, 85(4), 198-207.
- Mounts, N. S., & Steinberg, L. (1995). An ecological analysis of peer influence on adolescent grade point average and drug use. *Developmental Psychology*, 31(6), 915-922.
- Muthén, L., & Muthén, B. O. (2004). *Mplus users' guide* (3rd ed.). Los Angeles: Muthén & Muthén.
- Newcomb, A. F., & Bukowski, W. M. (1983). Social impact and social preference as determinants of children's peer group status. *Developmental Psychology*, 19(6), 856-867.
- Newsom, J. T., & Nishishiba, M. (2002). *Nonconvergence and sample bias in hierarchical linear modeling of dyadic data*. Unpublished manuscript, Portland State University, Portland, Oregon. Available from <http://www.upa.pdx.edu/IOA/newsom/papers.htm>
- Ollendick, T. H., Weist, M. D., Borden, M. C., & Greene, R. W. (1992). Sociometric status and academic, behavioral, and psychological adjustment: A five-year longitudinal study. *Journal of Consulting and Clinical Psychology*, 60(1), 80-87.
- Organisation for Economic Co-Operation and Development. (2005). Rationale for creating a global forum on education. In *Education Global Forum 2005*. Retrieved May 26, 2006, from [http://www.oecd.org/document/56/0,2340,en\\_21571361\\_35013845\\_35123640\\_1\\_1\\_1\\_1,00.html](http://www.oecd.org/document/56/0,2340,en_21571361_35013845_35123640_1_1_1_1,00.html)
- Parker, J. G., & Asher, S. R. (1993). Friendship and friendship quality in middle childhood: Links with peer group acceptance and feelings of loneliness and social dissatisfaction. *Developmental Psychology*, 29(4), 611-621.
- Parkhurst, J. T., & Hopmeyer, A. (1998). Sociometric popularity and peer-perceived popularity: Two distinct dimensions of peer status. *Journal of Early Adolescence*, 18(2), 125-144.
- Patterson, G. R., DeBaryshe, B. D., & Ramsey, E. (1989). A developmental perspective on antisocial behavior. *American Psychologist*, 44(2), 329-335.
- Patterson, G. R., Forgatch, M. S., Yoerger, K. L., & Stoolmiller, M. (1998). Variables that initiate and maintain an early-onset trajectory for juvenile offending. *Development and Psychopathology*, 10(3), 531-547.
- Patterson, G. R., Reid, J. B., & Dishion, T. J. (1992). *Antisocial boys*. Eugene, OR: Castalia.
- Pekarik, E. G., Prinz, R. J., Liebert, D. E., Weintraub, S., & Neale, J. M. (1976). The Pupil Evaluation Inventory: A sociometric technique for assessing children's social behavior. *Journal of Abnormal Child Psychology*, 4(1), 83-97.
- Pianta, R. C., & Stuhlman, M. W. (2004). Teacher-child relationships and children's success in the first years of school. *School Psychology Review*, 33(3), 444-458.
- Pittman, R. B. (1991). Social factors, enrollment in vocational/technical courses, and high school dropout rates. *Journal of Educational Research*, 84(5), 288-295.
- Rumberger, R. W. (1987). High school dropouts: A review of issues and evidence. *Review of Educational Research*, 57(2), 101-121.
- Rutter, M., & Maughan, B. (2002). School effectiveness findings 1979-2002. *Journal of School Psychology*, 40(6), 451-475.
- Schwartz, D., Gorman, A. H., Nakamoto, J., & McKay, T. (2006). Popularity, social acceptance, and aggression in adolescent peer groups: Links with academic performance and school attendance. *Developmental Psychology*, 42(6), 1116-1127.
- Schweinhart, L. J., Barnes, H. V., & Weikart, D. P. (1993). *Significant benefits: The High/Scope Perry Preschool Study through age 27*. Ypsilanti, MI: High/Scope Educational Research Foundation.
- Silver, R. B., & Eddy, J. M. (2006). Research-based prevention programs and practices for delivery in schools that decrease the risk of deviant peer influence. In K. A. Dodge, T. J. Dishion, & J. E. Lansford (Eds.), *Deviant peer influences in programs for youth: Problems and solutions* (pp. 253-277). New York: Guilford Press.
- Sullivan, H. S. (1953). *The interpersonal theory of psychiatry*. New York: W. W. Norton.
- Taylor, A. R. (1989). Predictors of peer rejection in early elementary grades: Roles of problem behavior, academic achievement, and teacher preference. *Journal of Clinical Child Psychology*, 18(4), 360-365.
- Tremblay, R. E., Desmarais-Gervais, L., Gagnon, C., & Charlebois, P. (1987). The Preschool Behaviour Questionnaire: Stability of its factor



- structure between cultures, sexes, ages and socioeconomic classes. *International Journal of Behavioral Development*, 10(4), 467–484.
- Tremblay, R. E., Vitaro, F., Gagnon, C., Piché, C., & Royer, N. (1992). A prosocial scale for the Preschool Behaviour Questionnaire: Concurrent and predictive correlates. *International Journal of Behavioral Development*, 15(2), 227–245.
- Vitaro, F., Brendgen, M., Larose, S., & Tremblay, R. E. (2005). Kindergarten disruptive behaviors, protective factors, and educational achievement by early adulthood. *Journal of Educational Psychology*, 97(4), 617–629.
- Vitaro, F., Larocque, D., Janosz, M., & Tremblay, R. E. (2001). Negative social experiences and dropping out of school. *Educational Psychology*, 21(4), 401–415.
- Vitaro, F., Pedersen, S., & Brendgen, M. (2007). Children's disruptiveness, peer rejection, friends' deviancy, and delinquent behaviors: A process-oriented approach. *Development and Psychopathology*, 19(2), 433–453.
- Vitaro, F., Tremblay, R. E., Gagnon, C., & Boivin, M. (1992). Peer rejection from kindergarten to Grade 2: Outcomes, correlates, and prediction. *Merrill-Palmer Quarterly*, 38(3), 382–400.
- Welsh, M., Parke, R. D., Widaman, K., & O'Neil, R. (2001). Linkages between children's social and academic competence: A longitudinal analysis. *Journal of School Psychology*, 39(6), 463–481.
- Wentzel, K. R. (2003). Sociometric status and adjustment in middle school: A longitudinal study. *Journal of Early Adolescence*, 23(1), 5–28.
- Wentzel, K. R., Barry, C. M., & Caldwell, K. A. (2004). Friendships in middle school: Influences on motivation and school adjustment. *Journal of Educational Psychology*, 96(2), 195–203.
- Woodward, L. J., & Fergusson, D. M. (2000). Childhood peer relationship problems and later risks of educational under-achievement and unemployment. *Journal of Child Psychology and Psychiatry*, 41(2), 191–201.

Received October 19, 2006

Revision received July 10, 2007

Accepted July 12, 2007 ■

### E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at <http://notify.apa.org/> and you will be notified by e-mail when issues of interest to you become available!

# Control, Motivation, Affect, and Strategic Self-Regulation in the College Classroom: A Multidimensional Phenomenon

Duane F. Shell  
University of Nebraska—Lincoln

Jenefer Husman  
Arizona State University

This study of 397 undergraduate students examined relations between self-reported control, goal orientation, future time perspective, affect, and strategic self-regulation. Five patterns were found in three canonical dimensions. The high end of bipolar Dimension 1 linked high self-regulated strategy use and study effort to high self-efficacy, outcome expectancy, and effort causal attribution; high mastery and performance approach and low work avoidance goal orientations; and positive affect. The low end of Dimension 1 linked low strategy use and effort to low self-efficacy, outcome expectancy, and effort causal attribution; high work avoidance goal orientation; and low affect. The high end of bipolar Dimension 2 linked knowledge-building strategies, but not active self-regulation or study effort, to high self-efficacy, outcome expectancy for learning but not grades, and affect causal attribution; high mastery goal orientation; and positive affect. The low end of Dimension 2 linked surface learning, consisting of active self-regulation and study effort but not personal knowledge building, to high effort causal attribution but low self-efficacy and outcome expectancy. Unipolar Dimension 3 linked learned helplessness to high outcome expectancy and external causal attribution but low self-efficacy; high work avoidance goal orientation; and high negative affect and anxiety.

*Keywords:* self-regulation, control beliefs, motivation, strategic learning, affect

The role of motivation in fostering students' self-regulated, strategic learning and academic performance has become an increasingly important topic in cognitive and educational psychology (e.g., Bandura, 1997, 2001; Boekaerts, Pintrich, & Zeidner, 2000; Pintrich, 2003; Pintrich & Schunk, 2002; Wigfield & Eccles, 2002; Zimmerman & Schunk, 2001). Current studies of motivation in academic achievement and self-regulated learning have been framed primarily within the broad context of achievement motivation as it has been extended from early conceptions by theorists such as Atkinson (1964) and Rotter (1966). In this tradition, personal beliefs are seen as one of the primary mechanisms affecting the motivation to engage and persist in self-regulatory and achievement-related behaviors.

Among the many beliefs and constructs identified within the motivational literature (see Pintrich, 2003), beliefs about control are some of the oldest and most examined (Pintrich, 2003; Skinner, 1996; Weisz & Stipek, 1982). Control beliefs concern the extent to which persons believe they can control or influence their environment to attain desired outcomes (e.g., Skinner, 1996). As well as constituting a distinct theoretical field, beliefs about control are thought to interact with other motivational constructs, such as goal

orientation and affect (Skinner, 1996; Pekrun, Elliot, & Maier, 2006; Walls & Little, 2005), and are central to many comprehensive theories of motivation and self-regulation, including social-cognitive theory (Bandura, 1997, 2001; Zimmerman, 2004) and self-determination theory (Reeve, Deci, & Ryan, 2004; Ryan & Deci, 2000).

Although control beliefs have had a prominent place in formulations of motivation (see Pintrich, 2003), there is no single control theory. Rather, control theories have provided an integrative framework for examining the overall motivational influence of a number of distinct constructs. As discussed by Weisz and Stipek (1982) and Skinner (1996), perceptions of control result from beliefs that behaviors and outcomes are contingently connected and that one possesses the competency or agency to successfully carry out those behaviors.

Competency or agent-means beliefs have been primarily studied as self-efficacy (Bandura, 1997, 2001; Pajares, 1996; Schunk & Pajares, 2004), although other similar competency formulations have appeared in the control literature (see Skinner, 1996, for a comprehensive review). Two types of contingency beliefs have been widely studied. Response–outcome contingencies, reflecting the perceived association between successful or unsuccessful actions and the attainment of outcomes or goals, have been studied as specific outcome expectancies for individual behaviors or domains (Bandura, 1997, 2001; Shell, Colvin, & Bruning, 1995; Shell, Murphy, & Bruning, 1989) and as locus of control, representing the generalized expectancy that outcomes are (internal) or are not (external) contingent on personal action (Chapman, Skinner, & Baltes, 1990; Heckhausen & Schulz, 1995; Rotter, 1966; Skinner, Chapman, & Baltes, 1988; Stipek & Weisz, 1981). Means–ends contingencies (Chapman et al., 1990; Heckhausen & Schulz, 1995; Skinner et al., 1988), reflecting perceived relations

---

Duane F. Shell, Department of Educational Psychology, University of Nebraska—Lincoln; Jenefer Husman, Division of Psychology in Education, Arizona State University.

This research was partially supported by U.S. Department of Education Secretary's Fund for Innovation in Education Grant R215D30195. An earlier version of this article was presented at the Annual Meeting of the American Educational Research Association, Chicago, March 1997.

Correspondence concerning this article should be addressed to Duane F. Shell, Department of Educational Psychology, 114 TEAC, University of Nebraska—Lincoln, Lincoln, NE 68588-0345. E-mail: dshell2@unl.edu



between possible causal influences and the production of successful or unsuccessful action, have been studied primarily within the causal attribution field (Weiner, 1986, 2004). Skinner and her colleagues (Chapman et al., 1990; Skinner, 1996; Skinner et al., 1988) have proposed an additional aspect of control consisting of agent–ends relations that reflect general feelings of control, which are evidenced by expectancy of success (see also Wigfield & Eccles, 2002).

### Control as a Multivariate Multidimensional Construct

Each of the constituent beliefs involved in producing perceptions of control has its own independent motivational influence. Studies have found that self-efficacy (Bandura, 1997, 2001; Pajares, 1996; Schunk & Pajares, 2004), outcome expectancies (Stipek & Weisz, 1981; Shell et al., 1995, 1989; Wigfield & Eccles, 2002), causal attributions (Weiner, 1986, 2004), and success expectancy (Wigfield & Eccles, 2002) all influence effort, choice, strategy use, self-regulation, and achievement in academic domains (see Pajares, 1996; Pintrich, 2003; Stipek & Weisz, 1981; and Wigfield & Eccles, 2002, for reviews). Control formulations, however, have sought to provide a framework for analyzing how these components interact to produce feelings of control. In control theories, control is a multivariate construct in which the whole is greater than the sum of its parts. Feelings of control arise from the overall pattern of agent–means relations (self-efficacy), response–outcome contingencies (outcome expectancies), means–ends contingencies (causal attributions), and agent–ends relations (expectancy of success) rather than from the level of any specific constituent control belief.

Because control is evidenced by the joint, simultaneous effects of the component control beliefs, control could potentially be a higher level, superordinate latent psychological construct that exists in addition to the specific component beliefs. As persons have beliefs about self-efficacy, outcome expectancy, and causal attributions, they might possess a cognitive belief that represents control. This interpretation, however, does not appear to be supported empirically. Factor analyses of component control beliefs (self-efficacy, outcome expectancies, causal attributions, and expectancy of success) have found that these are not reducible to a single latent dimension (Shell & Husman, 2001; Skinner, 1996; Skinner et al., 1988). Factor structures basically confirm that the component beliefs are distinct from one another, with each typically forming its own factorial dimension. As a result, current formulations view control as psychologically multidimensional (e.g., Chapman et al., 1990; Ford & Thompson, 1985; Heckhausen & Schulz, 1995; Skinner, 1996; Skinner et al., 1988; Walls & Little, 2005; Weisz & Stipek, 1982).

Because control does not appear to be a distinct psychological latent construct, feelings of control likely emerge during the course of thought and activity. Control formulations (e.g., Skinner, 1996; Walls & Little, 2005) view persons as active, self-regulating, goal-directed organisms. Feelings of control are intertwined with this active self-regulation. In school, elementary students' current and future control perceptions have been found to vary daily across assignments and in response to grades (Schmitz & Skinner, 1993). Shell et al. (1995, 1989) identified a canonical dimension linking an emergent pattern of control beliefs to reading and writing achievement. These findings suggest that patterns of control that

do not exist within control beliefs themselves can be found within the relations between control beliefs and other variables.

Weisz and Stipek (1982) argued that competency beliefs, consisting of self-efficacy and expectancy of success, and contingency beliefs, consisting of outcome expectancies, locus of control, and causal attributions, had different motivational consequences (see similar arguments by Bandura, 1997, about self-efficacy and outcome expectancy) and hence would be related differently to other variables. Shell and Husman (2001) found empirical support for competency and contingency dimensions of control in canonical correlations examining the relations between control beliefs and academic outcomes. They found separate canonical dimensions associating competency beliefs (self-efficacy) with grades and contingency beliefs (causal attributions and locus of control) with study time and effort. They also found that the dimensions of control in the canonical correlations could not be replicated by factor analysis of the control beliefs themselves. These findings suggest that there are multiple dimensions of control distinct from the multiple factorial dimensions of the component control beliefs.

### The Present Study

Our goal in this study was to further clarify the multidimensional nature of control as expressed in the relationships between control and academic effort, strategy use, and self-regulation. Although in recent theoretical work researchers have expressed a need for systematic, multidimensional characterizations of control within the broader context of motivation (Little, 1998; Skinner, 1996; Skinner et al., 1988), Walls and Little (2005) have noted a lack of empirical studies addressing linkages between control and other motivational constructs. Our first objective was to address this lack of empirical research by examining control beliefs in conjunction with a broader set of motivational constructs.

### Motivational Constructs

*Goal orientation.* Control formulations (e.g., Skinner, 1996; Walls & Little, 2005) hold that control is expressed in the context of goal-directed behavior. Therefore, the exercise of control is affected by a persons' goal orientation. Contemporary achievement goal theory (e.g., Pintrich, 2000a, 2000b) has focused on a distinction between goal orientations for mastery or learning and those for performance. Considerable research has linked these different goal orientations to specific control beliefs. Mastery goal orientation has been linked to more positive self-efficacy and task values, with these all associated with higher levels of strategic self-regulation (e.g., Wolters & Yu, 1996). Self-efficacy beliefs also have been linked to differing self-regulatory consequences of performance approach goal orientation (e.g., Midgley, Kaplan, & Middleton, 2001). Recent research has begun to link dysfunctional self-regulatory strategies, such as self-handicapping and defensive pessimism, to both performance (positive) and mastery (negative) goal orientations in conjunction with external attribution for success, ability attribution for failure, and uncertain control (Martin, Marsh, & Debus, 2003). This previous research strongly suggested that control beliefs and goal orientation are highly interconnected in their influences on students' achievement and self-regulation.

In this study, our objective was to move beyond looking at individual control beliefs and examine how dimensions of control,



such as competency and contingency, which have been found in the relations between control and aspects of self-regulation and achievement (e.g., Shell & Husman, 2001), are linked to students' goal orientation. The constellation of self-efficacy, causal attributions, and outcome expectancies supported by mastery goal orientation may be different than that supported by performance goal orientation. We were also interested in expanding the examination of goal orientation to include work avoidance goal orientation (Ames, 1992; Wolters, 2003), because the control implications of work avoidance goal orientation have not been studied.

*Future time perspective (FTP).* Formulations of control beliefs have long contained references to the link between control and the future. Weiner (1986, 2004) noted that the degree to which causal attributions are seen as continuing into the future plays a major role in their motivational force. Bandura (2001) emphasized that cognitive belief systems exist to allow for forethought and extension of thinking into the future. Control theorists (Skinner, 1996; Schmitz & Skinner, 1993) have presumed that control beliefs influence interpretations of current achievement that subsequently affect future control beliefs and performance.

A person's conceptualization of the future and connection to that future have been referred to as *future time perspective* (FTP; Husman & Lens, 1999; Lens, 1987). Two aspects of FTP that have been found to be of particular relevance to academic achievement are defined as follows: (a) value or valence, defined as the disposition to ascribe high value to goals in the future relative to goals in the present, and (b) connectedness, defined as the disposition to anticipate, in the present, the long-term consequences of a potential action (see De Volder & Lens, 1982; Husman, Derryberry, Crowson, & Lomax 2004; Husman & Lens, 1999; Shell & Husman, 2001). Shell and Husman (2001) found that these FTP constructs were jointly related, with control beliefs, to achievement and studying, with higher connectedness and valence supporting both competency and contingency control dimensions. Miller, Greene, Montalvo, Ravindran, and Nichols (1996) also have found that FTP interacted with goals and perceived ability to predict academic engagement.

These studies suggest that FTP plays a potential role in supporting the influence of control on self-regulation and achievement. Husman and Lens (1999) have argued that FTP can make distal outcomes more salient in current thought; therefore, we would expect students with higher FTP to gain more motivational influence from expectancies for longer range outcomes, potentially increasing the motivational force of outcome expectancies. We also would expect students with higher FTP to have more motivation from attribution to causes that have future stability. To continue the examination of these linkages between control and FTP, we included FTP measures for valence and connectedness in the study.

*Affect.* Feelings of control are thought to produce positive affective responses (e.g., Pekrun, Goetz, Titz, & Perry, 2002; Reeve et al., 2004). Weiner (1986, 2004) has emphasized the different positive and negative affective reactions that result from different attributional patterns. Self-worth theory (Covington, 2004) has also addressed the affective components that underlie why students act and attribute causality in ways that protect positive feelings of worth. Low self-efficacy has been linked to test anxiety (Pajares, 1996).

In the control-value theory of achievement emotion, Pekrun (Pekrun et al., 2006, 2002) has proposed that emotional appraisals are predominantly influenced by control, specifically competence or self-efficacy perceptions, causal attributions, and subjective value, or what are defined here as outcome expectancies. Feelings of control produce positive emotional reactions, whereas feelings of a lack of control produce negative emotional reactions. In addition, Pekrun et al. (2006) have linked control and emotions to goal orientations, arguing that mastery and performance goal orientations influence control perceptions and subsequent emotions. Although Pekrun et al. empirically confirmed predictions for goal-emotion relationships, they did not include explicit measures of control.

Following Pekrun's logic (Pekrun et al., 2006, 2002), we wanted to extend the examination of how emotion, control, and goals affect strategic self-regulation in classrooms. We expected that emotions would be significant components of any control dimensions in ways predicted by Pekrun.

### *Differential Control Beliefs for Different Academic Outcomes*

Most control research has used general measures of academic control (e.g., the Control, Agency, and Means-Ends Interview; Skinner et al., 1988). These measures assume that students have the same control beliefs across all academic outcomes. However, students' beliefs about response-outcome contingencies (outcome expectancies), means-ends contingencies (causal attributions), agent-means competency (self-efficacy), and agent-ends relations (expectancy of success) might be different for learning course material versus achieving an outcome such as high grades. Distinctions between more intrinsic outcomes resulting from learning and extrinsic outcomes resulting from rewards or attaining external criteria are central to many motivational theories, such as self-determination theory (Reeve et al., 2004; Ryan & Deci, 2000), but have not been extensively examined in the control literature.

A second objective, therefore, was to examine whether students had different control beliefs for learning versus achievement outcomes. We assessed causal attributions and outcome expectancies separately for learning course material and obtaining a high grade. For agent-means and agent-ends relations, we assessed both self-efficacy for executing effective strategies for learning course material and expectancy for achieving a high grade.

Because control is expressed within the context of goal orientation (e.g., Skinner, 1996; Walls & Little, 2005), we expected that any differences between students' control beliefs for learning versus achievement outcomes would be expressed in conjunction with differential goal orientation. We expected that patterns of control beliefs for learning would be more strongly linked to mastery goal orientation and that those for achievement would be more strongly linked to performance approach goal orientation.

### *Strategic, Self-Regulatory Variables*

A final objective was to examine control within a broader self-regulatory context than has been previously studied. Shell and Husman (2001) found that competency and contingency beliefs were differentially associated with achievement and studying. They noted that the association they found between competency



and achievement was likely mediated by student learning strategies; however, Shell and Husman did not have measures of cognitive or self-regulatory strategies in their study. Also, Shell and Husman speculated that contingency control beliefs perhaps motivated more general levels of effort and persistence rather than specific strategy use. To examine these questions, we used a more comprehensive array of self-regulation, strategy, and effort measures.

We have included strategic self-regulatory constructs reflecting two somewhat different theoretical and research traditions. The first tradition comes from information processing or, more recently, self-regulation theory (see discussion in Pintrich, 2004). Evolving from early work on study strategies (Weinstein & Mayer, 1986) and metacognition (Brown, Bransford, Ferrara, & Campione, 1983), research in these areas has led to a model of strategic and self-regulated learning that Pressley, Borkowski, and Schneider (1987) have termed *the good strategy user*. Good strategy users are metacognitively active through planning, monitoring, and evaluating their learning and strategy use. They also use a variety of strategies to help them understand, remember, and organize the information they are learning (Pressley et al., 1987; Weinstein, Husman, & Dierking, 2000; Weinstein & Mayer, 1986; Zimmerman & Martinez-Pons, 1988; Zimmerman & Schunk, 2001).

The second tradition has evolved within the more constructivist, knowledge-building approach to learning proposed by Bereiter and Scardamalia (1989; Chan, Burtis, & Bereiter, 1997). Central to the knowledge-building approach is the idea that meaningful learning involves the production of knowledge rather than the reproduction of knowledge. This knowledge building is accomplished by an in-depth study of a topic that goes beyond simple factual or recall learning and is supported by active engagement in classrooms as evidenced by question asking (Scardamalia & Bereiter, 1992).

In addition, because research has begun to examine more dysfunctional self-regulatory strategies (e.g., Martin et al., 2003; Wolters, 2003), we also incorporated assessment of negative strategies similar to the lack of regulation identified by Vermunt and Vermetten (2004; see also Dweck & Leggett, 1988; Zimmerman & Martinez-Pons, 1988). By examining a broad range of potential strategic self-regulatory behaviors, we hoped to better capture the extent and complexity of the relations between control, academic effort, strategy use, and self-regulation.

### Research Questions

Specific research questions for the study were as follows:

1. Are there different dimensions of control associated with different patterns of strategic self-regulation? Specifically, we hypothesized that competency and contingency beliefs would form distinct control dimensions that are associated with different patterns of strategic self-regulation.
2. Are different dimensions of control connected to different goal orientations? Specifically, we hypothesized that control beliefs would form unique dimensions differentiated by mastery, performance approach, and work avoidance goal orientations.
3. How do control beliefs for learning versus achievement outcomes differ in their influence on strategic self-regulation, and are these differences connected to mastery and performance approach goal orientations? Specifically, we hypothesized that control beliefs for learning and grades would be more strongly con-

nected to mastery and performance approach goal orientations, respectively.

4. How are dimensions of control beliefs connected to FTP? Specifically, we hypothesized that higher perceived control would be connected to higher FTP.

5. How are dimensions of control connected to different patterns of affect? Consistent with Pekrun et al. (2006), we hypothesized that higher perceived control would be connected to more positive affect and lower anxiety.

## Method

### Participants

Participants were 397 undergraduates (150 men, 242 women, 5 unknown) at a large Southwestern university who volunteered to participate after being recruited from multiple sections of two upper division educational psychology courses. Participants received course research participation credit for their participation in the study. Students who did not wish to participate in the research were offered the opportunity to complete an alternative assignment for credit.

### Control and Motivation Measures

With the exception of the locus-of-control and FTP measures, all instruments asked students about beliefs concerning the specific class from which they were recruited.

**Self-efficacy.** Self-efficacy for self-regulation and use of learning strategies in the class was assessed by asking participants to rate on a scale of 0 (*no chance*) to 100 (*complete certainty*) their confidence in successfully doing a series of 22 self-regulatory and study strategies (e.g., "Take effective notes over course lectures," "Manage time well enough to have ample study time for the class," "Study effectively for the course exams"). Principal components analysis of the items verified that the items formed a single factor. The self-efficacy score was computed as the mean score of the scale items. Item and reliability analysis indicated acceptable scale properties, with item-total correlations ranging from .46 to .78 and Coefficient  $\alpha = .94$ .

**Expectancy for success.** Expectancy for success (e.g., Wigfield & Eccles, 2002) was assessed by asking participants to rate on a scale of 0 (*no chance*) to 100 (*complete certainty*) their confidence in achieving a grade of A in the course. This type of measure is similar to measures of self-efficacy for grades (e.g., Zimmerman & Bandura, 1994).

**Causal attributions.** Attributions were assessed by asking participants to rate on a 4-point Likert scale from 1 (*very unimportant*) to 4 (*very important*) how important they believed each of 10 possible causes was for producing success and failure in the class from which they were recruited. Attributions were assessed for both an achievement outcome (achieving a high grade) and a learning outcome (learning the course material). Separate principal components analyses of the items for success and failure indicated no differences in the attribution factor structures, so these were combined and factored as a single set. Separate principal components analyses of grade and learning attributions also found no differences in factor structure. However, because grade and learn-



ing attributions still might have different relations with strategic self-regulatory variables, these were retained as separate scales.

For both scales, five factors were identified: (a) strategic learning/effort, consisting of attributions to normal effort, extra effort, and good study skills; (b) ability, consisting of attributions to intelligence and prior knowledge; (c) affect, consisting of attributions to interest and enjoyment; (d) external causes, consisting of attributions to luck and task difficulty; and (e) help, consisting of attributions to obtaining help from friends or teachers. Although these factors do not exactly correspond to Weiner's (1986, 2004) three proposed dimensions of control, they are consistent with those found in previous studies (e.g., Hamilton & Akhter, 2002; Vispoel & Austin, 1995; Weiner, 1986), although luck and task difficulty sometimes form individual factors and sometimes, as we found, combine into a more global external factor.

Scale scores for each of the five factors were produced by taking the mean of the items in each factor. Item and reliability analysis indicated acceptable scale properties for each attribution scale. Item-scale correlations for all but four items with their respective scales were above .40, with the remaining four items between .30 and .37. Coefficient alpha reliability estimates for the grade attribution scales were strategic learning/effort = .78, ability = .69, interest/enjoyment = .81, external causes = .65, and help = .77. Coefficient alpha reliability estimates for the learning attribution scales were strategic learning/effort = .80, ability = .68, interest/enjoyment = .82, external causes = .74, and help = .81.

**Outcome expectancy.** Outcome expectancy was measured by two scales adapted for the classroom situation from the scales used by Shell et al. (1989). The first instrument assessed outcome expectancies for an achievement outcome. Students were asked to rate the importance of having a high grade in their class for successfully achieving 15 life goals drawn from the domains of employment, social activities, family life, education, and general citizenship. The second instrument assessed outcome expectancies for a learning outcome. Students were asked to rate the importance of learning the class content for achieving the same 15 life goals. Both instruments were answered on a 4-point, Likert-type scale ranging from 1 (*very unimportant*) to 4 (*very important*). Scale scores were computed as the mean score of the scale items. Item and reliability analysis indicated acceptable scale properties, with item-scale correlations ranging from .34 to .69 for the grade expectancy scale and .34 to .67 for the learning expectancy scale. Coefficient alpha estimates for the grade expectancy and learning expectancy scales were .89 and .88, respectively.

**Locus of control.** Locus of control was measured using the Rotter Internal-External Locus of Control Scale (Rotter, 1966). The instrument is a 29-item questionnaire with 23 assessment items and 6 filler items. Each assessment item offered a choice between an internal response and an external response, and participants were asked to indicate which of the two choices they most strongly believed to be true. The instrument is commonly scored in the external direction; however, to make the direction of the scores consistent with the scales for causal attribution measures of internality and externality, the scoring was done by summing the internal responses so that higher scores indicated a more internal locus of control.

**Goal orientation.** Students' personal goal orientations for the class were measured using a scale based on Dweck's formulation (e.g., Dweck & Leggett, 1988) adapted from Schraw, Horn,

Thorndike-Christ, and Bruning (1995). In response to the following stem: "Students differ in what they want to get out of the courses they take. Use the scale given to rate how important achieving each of the following is for you," students rated how important achieving each goal was in the class on a 4-point Likert scale ranging from 1 (*very unimportant*) to 4 (*very important*).

Five items reflected mastery or learning goal orientation: (a) "Learning new knowledge or skills in the class just for the sake of learning them," (b) "Really understanding the course material," (c) "Being challenged by course assignments," (d) "Feeling satisfied that you got what you wanted from the course," and (e) "Knowing more than you did previously about the course topics." Five items reflected performance approach goal orientation (e.g., Wolters, 2003): (a) "Doing better than the other students in the class on tests and assignments," (b) "Proving to other people that you are a good student," (c) "Remembering enough from the class to impress other people," (d) "Getting the highest grade in the class," and (e) "Impressing the instructor with your performance."

Because the course from which most students were recruited was a highly popular elective class in which achievement was not directly related to students' majors, we did not feel that performance avoidance goal orientation (Pintrich, 2000a, 2000b) was likely to be very applicable. However, because the course was often taken because it was seen as "easy," we did feel that goals reflecting work avoidance orientation (Ames, 1992; Wolters, 2003) might be applicable. Therefore, we added three items reflecting work avoidance goal orientation: (a) "Not having to work too hard in the class," (b) "Getting a passing grade with as little studying as possible," and (c) "Getting through the course with the least amount of time and effort."

A principal components analysis verified the three distinct goal orientation scales. Item and reliability analysis indicated acceptable scale properties, with item-scale correlations ranging from .37 to .58 for the learning goal orientation scale, from .42 to .64 for the performance approach goal orientation scale, and from .73 to .87 for the work avoidance goal orientation scale. Coefficient alpha estimates for the learning, performance approach, and work avoidance goal orientation scales were .72, .79, and .78, respectively. Scale scores were computed as the mean of the items in each goal orientation scale.

**FTP.** FTP was measured with the Future Time Perspective (FTP) Scale (Shell & Husman, 2001). The FTP Scale is a 25-item instrument containing two subscales: Connectedness and Valence. The FTP Connectedness scale consists of 16 questions that assess the contingent or instrumental relationship between current behavior and future goal attainment. The FTP Valence scale consists of 9 questions that assess the relative importance of attaining immediate versus long-range future outcomes. Participants are asked to indicate their agreement with each question using a 5-point Likert scale ranging from 1 (*strongly disagree*) to 5 (*strongly agree*). Scale scores are computed as the mean of the items in each scale. The complete scale with instrument validation psychometrics can be found in Shell and Husman (2001). For the sample in the present study, the coefficient alpha reliability estimates for the FTP Connectedness and Valence subscales were .86 and .83, respectively.

**Course affect.** Affect was measured by the Positive and Negative Affect Scale (Watson, Clark, & Tellegen, 1988), a 22-item scale that asks students to rate the frequency with which they



experience 22 emotions (11 positive and 11 negative) in their class on a 5-point scale ranging from 1 (*a few times or none*) to 5 (*most of the time, 80%–100% of the time*). Positive emotions are interested, excited, attentive, active, enthusiastic, proud, alert, capable, in control, inspired, and determined. Negative emotions are scared, out of control, irritable, ashamed, nervous, distressed, upset, guilty, hostile, frustrated, and afraid. A principal components analysis verified the existence of separate positive and negative scales. Item and reliability analysis indicated acceptable scale properties, with item–scale correlations ranging from .38 to .76 for the Positive Affect scale and from .49 to .72 for the Negative Affect scale. Coefficient alpha estimates for the Positive and Negative Affect scales were .88 and .88, respectively. Scale scores were computed as the mean of the items in each scale.

**Course anxiety.** The course anxiety scale was developed as an extension of test anxiety measures in instruments such as the Motivated Strategies for Learning Questionnaire (MSLQ; Pintrich, Smith, Garcia, & McKeachie, 1993). Assessment of anxiety was extended from testing situations to a broader set of 15 class-related situations (e.g., “When I take tests, I think about the consequences of failing”; “When I’m in class, I worry about being called on and not knowing the answer”; “I fear having to do tests, papers, and other graded assignments”; “When I do assignments, I worry about how well I will do”; “When I take a test, I get so worried that my mind just goes blank”). Students were asked to rate the frequency on a 5-point Likert-type scale ranging from 1 (*almost never*) to 5 (*almost always*) that they felt concern, anxiety, or worry about each of the 15 situations. Principal components analysis of the items verified that the items formed a single factor. Item and reliability analysis indicated acceptable scale properties, with item–total correlations ranging from .33 to .71 and coefficient  $\alpha = .90$ . Scale scores were computed as the mean of the 15 items.

### Strategic Self-Regulatory Measures

Strategic self-regulation was assessed with the Student Perceptions of Classroom Knowledge-Building Scale (SPOCK; Shell et al., 2005). The SPOCK was developed to extend the assessment of student self-regulated, strategic learning beyond current measures such as the MSLQ (Pintrich et al., 1993) and the Learning and Study Strategies Inventory (LASSI; Weinstein, Zimmermann, & Palmer, 1988) by including strategies, question asking, and behaviors based on the knowledge-building, question-asking, and intentional learning models of Bereiter and Scardamalia (1989; Chan et al., 1997) and negative aspects of student regulation as portrayed in descriptions of learned helplessness (e.g., Dweck & Leggett, 1988) and in the lack-of-regulation strategy of Vermunt and Vermetten (2004). The SPOCK also contains scales that were not used in this study that assess students’ perceptions of collaboration and teacher directedness in the classroom.

The instrument asks students about strategic self-regulatory behavior within a specific course. Students were asked to respond only for the class from which they were recruited. All questions were answered on a 5-point Likert scale. Scale points were described as 1 (*almost never—occurs on a very rare occasion or not at all*), 2 (*seldom—does not occur often; occurs about one quarter of the time*), 3 (*sometimes—occurs about half of the time*), 4 (*often—occurs frequently; occurs about three quarters of the time*), and 5 (*almost always—usually or always occurs; on a rare*

*occasion it may not occur*). The SPOCK measures four aspects of students’ perceptions of their own strategic self-regulation: self-regulated strategy use, knowledge building, question asking (both high level and low level), and lack of regulation.

The Self-Regulated Strategy Use scale (8 items) assesses the extent of student planning, goal setting, monitoring, and evaluation of studying and learning: (a) “In this class, I try to determine the best approach for studying each assignment”; (b) “In this class, I try to monitor my progress when I study”; (c) “In this class, I make plans for how I will study”; (d) “In this class, I use different ways to organize my thoughts, such as diagrams, charts, timetables, etc.”; (e) “In this class, I check myself to see how well I am understanding what I am studying”; (f) “In this class, I take notes and jot down questions when I am reading the class materials”; (g) “In this class, I focus on understanding the important ideas in what I am reading or studying”; and (h) “In this class, I set goals for myself which I try to accomplish.” These items assess strategic behaviors and study strategies typically associated with models of strategic self-regulation (e.g., Pintrich, 2004; Weinstein & Mayer, 1986) and what Pressley et al. (1987) have termed the good strategy user.

The Knowledge-Building scale (8 items) assesses the extent of student exploration and interconnection of knowledge from the class on the basis of the knowledge-building and intentional learning models of Bereiter and Scardamalia (1989; Chan et al., 1997). Questions in this scale focus on going beyond the given material and on tying the information being learned to other courses and existing knowledge: (a) “Whenever I learn something new in this class, I try to tie it to other facts and ideas that I already know”; (b) “As I study the topics in this class, I try to think about how they relate to the topics I am studying in other classes”; (c) “In this class, I set goals based on things I really want to learn”; (d) “In this class, I ask questions that allow me to explore topics that interest me”; (e) “As I study topics in other classes, I try to think about how they relate to the topics I am studying in this class”; (f) “In this class, I do assignments primarily to learn something new”; (g) “In this class, I ask questions that can only be answered by exploring new information”; and (h) “As I study a topic in this class, I try to consider how the topic relates to other things I know about.”

Two scales assess the extent of question asking in class (see Scardamalia & Bereiter, 1992). The High-Level Question-Asking scale (4 items) assesses the extent to which students ask questions that extend or expand on the basic information being provided in the class: (a) “In this class, I ask questions in order to help me learn new things”; (b) “In this class, I ask questions about things I am curious about”; (c) “In this class, I ask questions to help me know more about the topics we are covering in class”; and (d) “In this class, I ask questions to help me better understand the things I am trying to learn.” The Low-Level Question-Asking scale (4 items) assesses the extent to which students ask questions to obtain or clarify basic course information: (a) “In this class, I ask questions so that I can be sure I know the right answers for tests”; (b) “In this class, I ask questions to be clear about what the instructor wants me to learn”; (c) “In this class, I ask questions to help me prepare for tests”; and (d) “In this class, I ask questions so that I can find out what information the instructor thinks is important.”

The Lack-of-Regulation scale (4 items) assesses students’ lack of understanding of how to study and need for assistance and guidance in studying: (a) “In this class, when I get stuck or



confused about my schoolwork, I need someone else to figure out what I need to do"; (b) "In this class, I have trouble figuring out how to approach studying"; (c) "In this class, I rely on the instructor to tell me what to do"; and (d) "In this class, when I don't know an answer to a question, I try to hide the fact that I don't know." This scale assessed behaviors similar to those in the lack-of-regulation orientation identified by Vermunt and Vermetten (2004).

Scale scores were computed as the mean score of the scale items. Coefficient alpha reliability estimates for the Self-Regulated Strategy, Knowledge-Building, High-Level Question-Asking, Low-Level Question-Asking, and Lack-of-Regulation scales were .81, .84, .92, .91, and .48, respectively.<sup>1</sup> With the exception of lack of regulation, these reliability estimates for the SPOCK are consistent with those obtained for similar instruments, such as the MSLQ (Pintrich et al., 1993), the LASSI (Weinstein et al., 1988), and the Study Process Questionnaire (Biggs, Kember, & Leung, 2001).

Hamman (1998) found that SPOCK scores for the Self-Regulated Strategies, Knowledge-Building, and Lack-of-Regulation scales were correlated in expected ways with scores on comparable subscales of the LASSI (Weinstein et al., 1988). Shell et al. (2005) found that SPOCK scores, especially Lack-of-Regulation scale scores, predicted high school students' course grades.

In addition to the SPOCK, two measures of studying taken from Shell and Husman (2001) were used. Study time was assessed by asking participants to indicate the average number of hours per week they spent studying for the class from which they were recruited on a scale ranging from 1 to 7 representing 5-hr units from 1 (*less than 5 hours per week*) to 7 (*over 30 hours per week*). Perceived study effort was assessed by asking participants to indicate their perceptions of the effort they put forth studying for the class from which they were recruited relative to most other students in the class on a 5-point Likert scale with the following point designations: 1 (*I put forth much less effort studying*), 2 (*I put forth somewhat less effort studying*), 3 (*I put forth about the same effort studying*), 4 (*I put forth somewhat more effort studying*), and 5 (*I put forth much more effort studying*).

### Procedures

Data were collected over two semesters. In both semesters, participants in groups of 10 to 20 completed the measures in a single session proctored by one of the researchers or a graduate research assistant. In one semester, measures were administered in two packets, with students answering on Scantron mark sense sheets for one packet and on the packet itself for the other packet. In the other semester, measures were administered in two packets, with students answering on Scantron mark sense sheets but answering self-efficacy questions on their own form. Demographic questions about gender, age, and year in school were included in the packets used in both semesters.

### Data Analysis

We conducted canonical correlation analyses using the canonical correlation macro in SPSS for Windows Version 13.0. The strategic self-regulatory variable set consisted of SPOCK scores

along with study time and perceived study effort. The control and motivation variable set consisted of self-efficacy, expectancy for success, causal attribution, outcome expectancy, locus of control, goal orientation, FTP, course affect, and course anxiety variables.

## Results

### Descriptive Statistics

The means, standard deviations, and zero-order correlations between all variables are provided in Table 1. The average profile of students' control, motivation, and strategic self-regulation as portrayed in the mean scores fit the good strategy user described by Pressley et al. (1987; see also Weinstein et al., 2000; Zimmerman & Schunk, 2001), which would be expected of primarily junior and senior students taking an upper level course. Students reported about average levels of self-regulated strategy use, somewhat infrequent knowledge building, and infrequent question asking and lack of regulation. They reported studying for the course about 5–10 hr per week with somewhat of a below-average effort. These levels of studying and self-regulated strategic behavior seem consistent with what would be expected for students in a general elective course. Students expressed generally high levels of self-efficacy for using learning strategies and high expectancy for grades. They attributed success and failure most to effort/strategy use and affect (interest and enjoyment) for both learning and grades. Students had moderate levels of outcome expectancy for both learning and grades and neither a strong internal nor a strong external locus of control. These control patterns are consistent with typical patterns found among college and even younger student populations (e.g., Skinner, 1996). Students expressed higher mastery goal orientation than either performance approach or work avoidance orientation, which were similar to each other. They had high FTP connectedness and middle-range FTP valence, suggesting a generally high overall FTP. They expressed generally high positive affect for the class and lower levels of negative affect and anxiety.

### Canonical Correlation Results

We first present the results of the canonical correlation analysis. We then address each of our research questions. Results of the

<sup>1</sup> The coefficient alpha estimate for lack of regulation is consistent with those found for the same four-item scale in two high school samples (.54 and .50) by Shell et al. (2005). Coefficient alpha represents the minimum possible scale reliability. True reliability cannot be lower but may be higher. Two scale conditions can depress alpha relative to true reliability (see Thorndike, 1997). The first is skewed item responses. Because few college students perform the behaviors assessed in the Lack-of-Regulation scale, item scores for three of the four items were heavily skewed, with 60%–80% of students answering 1 or 2 on the 5-point scale. The second is Spearman-Brown suppression. Coefficient alpha can underestimate reliability in short scales, especially if skewed items are present. A revised version of the SPOCK with a longer seven-item Lack-of-Regulation scale had an alpha of .63 with a high school sample (Shell et al., 2005) and .64 with a college sample (Shell, 2005). These later estimates suggest that for the current sample the coefficient alpha estimate is experiencing Spearman-Brown suppression, and true scale reliability is likely closer to the .64 attained with the longer scale.



Table 1  
Means, Standard Deviations, and Intercorrelations of Measured Variables

Variable	Strategic self-regulatory variables										Control and motivational variables										M	SD											
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20			21	22	23	24	25	26	27	28	29	30	
Strategic self-regulatory variables																																	
1. Self-regulated strategy use	—	.42	.32	.37	.13	.46	.31	.20	-.07	.37	.21	.01	-.01	.27	.34	.23	.06	-.01	.26	.08	.12	.16	.28	.23	-.38	.06	.14	.16	.07	.17	3.07	.80	
2. Knowledge building	—		.53	.50	.10	.10	.12	.25	.10	.09	.12	.19	.04	.13	.10	.13	.22	.04	.12	.42	.30	-.01	.48	.25	-.12	.00	.11	.41	-.02	.16	2.65	.79	
3. High-level question asking	—			.89	.07	.20	.10	.16	.03	.09	.01	.09	.03	.23	.12	.02	.11	-.05	.20	.25	.24	.07	.26	.19	-.25	-.02	.07	.22	-.00	.07	2.25	1.03	
4. Low-level question asking	—				.11	.22	.12	.12	.00	.14	.04	.09	.09	.30	.19	.08	.14	.02	.30	.25	.27	.08	.21	.22	-.20	-.01	.11	.27	.05	.15	2.35	1.06	
5. Lack of regulation	—					.03	.10	-.22	-.28	.05	.13	.05	.25	.22	.08	.11	-.00	.21	.18	.15	.15	-.02	-.06	.14	.08	-.10	.12	-.06	.32	.37	2.31	.67	
6. Study time	—						.41	.04	-.18	.25	.04	-.05	.02	.17	.23	.11	-.04	-.04	.11	.02	.04	-.02	.14	.06	-.33	-.01	.06	-.04	.14	.10	1.68	1.07	
7. Perceived study effort	—							.22	.03	.16	.03	.04	.03	.13	.17	.10	.02	.01	.11	.11	.14	-.06	.14	.15	-.28	-.03	.00	.11	.09	.12	2.73	.94	
Control and motivational variables																																	
8. Self-efficacy	—								.55	-.04	.03	-.07	-.17	-.13	-.04	.02	.00	-.15	-.06	.01	-.01	.07	.31	.18	-.19	.02	-.01	.34	-.35	-.19	82.94	11.99	
9. Expectancy for grade success	—									-.21	-.10	-.08	-.21	-.27	-.19	-.13	-.06	-.14	-.21	-.03	-.03	-.01	.12	.15	-.08	.04	-.05	.32	-.29	-.22	73.87	27.83	
10. Learning attributions strategy/effort	—										.26	.33	.22	.44	.75	.31	.35	.21	.39	.12	.22	.14	.20	.14	-.19	.25	.18	.01	.05	.13	3.16	.52	
11. Learning attributions ability	—											.20	.28	.28	.23	.77	.19	.33	.24	.18	.17	.02	.10	.13	.05	-.00	-.01	.01	.04	.11	2.57	.60	
12. Learning attributions interest/enjoyment	—													.31	.23	.30	.25	.63	.28	.21	.27	-.20	-.06	.17	.19	-.05	.13	-.07	.09	.29	2.33	.63	
13. Learning attributions external causes	—														.40	.26	.36	.32	.70	.24	.27	-.00	.10	.20	-.12	-.02	.03	.08	.16	.27	2.78	.77	
14. Learning attributions help	—																.38	.44	.29	.51	.24	.33	.09	.28	.19	-.18	.22	.18	.03	.05	.23	3.19	.52
15. Grade attributions strategy/effort	—																		.28	.26	.22	.23	-.00	.13	.16	.01	.05	.07	-.01	.08	.17	2.59	.62
16. Grade attributions ability	—																		.38	.29	.30	-.07	.31	.08	.08	.10	.07	.16	-.09	.13	3.21	.65	
17. Grade attributions interest/enjoyment	—																			.34	.21	.35	-.25	.04	.15	.28	.06	.12	-.01	.07	.31	2.57	.61
18. Grade attributions external causes	—																				.22	.31	-.05	.17	.22	-.12	.10	.08	.01	.13	.26	2.85	.80
19. Grade attributions help	—																																
20. Learning outcome expectancy	—																																
21. Grade outcome expectancy	—																																
22. Locus of control	—																																
23. Mastery goal orientation	—																																
24. Performance approach goal orientation	—																																
25. Work avoidance goal orientation	—																																
26. FTP connectedness	—																																
27. FTP valence	—																																
28. Positive class affect	—																																
29. Negative class affect	—																																
30. Course anxiety	—																																

Note. Correlations of .13 and above are significant at  $p < .01$ , two-tailed. FTP = future time perspective.

canonical correlation analysis are shown in Table 2. For each variable set, a canonical variate, which is a latent linear combination of the variables, similar to a factor in factor analysis, is created. The canonical correlation is the correlation between these canonical variates for the two variable sets. Of the seven possible canonical correlations (equal to the number of variables in the smallest set), five were statistically significant on the basis of sequential chi-square testing (see Stevens, 2002; Thompson, 1984). Examination of explained variance in the correlation ( $R_c^2$ ) and redundancy analysis, however, suggested that only the first three correlations were meaningful. Each of the first three correlations exceeded .50, with the set of control and motivation variables accounting for 28% to 45% of the variance in strategic self-regulatory variables. These results indicated meaningful prediction of strategic self-regulation from the control and motivation variables. No other canonical correlation was higher than .38. Analysis of redundancy variance found that the variates for the first three canonical correlations combined explained 61% of the variance in the strategic and self-regulatory variables. Variates for the remaining two significant correlations explained only 20% of the variance in the strategic and self-regulatory variables.

For each variable set, the canonical variate describes a latent dimension. This latent dimension portrays the combined influence of the variables in the set. Identifying what the latent dimension represents is done by interpreting which variables in the set are making meaningful contributions to the variate. Contribution is determined from two indicators: canonical coefficients and structure coefficients. The canonical coefficients are the weighting of each variable in the canonical discriminant function that defines the variate. They reflect the independent contributions of each variable to the variate with the other variables in the set partialled out, similar to beta weights in a regression equation. The structure coefficients (or loadings) are the correlation between participants' scores on an original variable and their canonical variate scores (scores produced by multiplying the scores for each variable by their respective canonical coefficients). They are the same as factor loadings in a factor analysis. If variables in the set are intercorrelated, the pattern of canonical coefficient weights can be arbitrary. Because of this, the structure coefficients are usually preferred for interpreting which variables are contributing to a canonical variate. However, the structure coefficients may not reflect all of the contributions if the variables are not highly intercorrelated. Also, if variables that are themselves intercorrelated have independent contributions to the canonical variate in opposite directions (one

positive and one negative), neither may be correlated very highly with the resulting linear combination (see Shell & Husman, 2001; Stevens, 2002; or Thompson, 1984, for more complete discussions of interpreting contributions to canonical variates). In general, we based our decisions on the structure coefficients. We considered a structure coefficient above .30 to indicate meaningful contributions. However, if a variable had a high canonical coefficient relative to other variables, we also considered it to be a contributor to the variate, even if its structure coefficient was less than .30. The canonical coefficients and structure coefficients for each of the three canonical correlations are provided in Table 3.

The canonical and structure coefficients can identify which variables make a contribution to the canonical variates. Like a factor in factor analysis, however, the identified pattern of variable contribution is meaningful only in the context of the theory and literature of the specific constructs and behaviors being examined. Also, the canonical variates for each correlation define a dimensional continuum that can be unipolar or bipolar. In a unipolar dimension, movement along the continuum indicates change from low to high in a unitary entity, such as an increase in self-efficacy or an increase in math or reading ability. In a bipolar dimension, movement along the continuum indicates a shift in qualitatively distinct states, such as a change from introversion to extroversion. The low end of the dimension can be examined by changing the sign of the canonical and structure coefficients from positive to negative and vice versa. As with the interpretation of the meaning of a canonical variate, decisions about whether a canonical dimension is unipolar or bipolar are based on whether the ends of the continuum define patterns that are theoretically interpretable in relation to the literature for the variables in the dimension.

*Structure of the first canonical dimension.* At the high end of the first canonical dimension, the structure coefficients portrayed a pattern of higher self-regulated strategy use, knowledge building, high- and low-level question asking, study time, and study effort. This pattern of strategic self-regulation was associated with a pattern of structure coefficients portraying higher positive control beliefs (higher self-efficacy, higher causal attributions for both learning and grades to strategy use/effort and help and for grades to ability, and higher outcome expectancy for learning and grades), higher mastery and performance approach goal orientations and lower work avoidance goal orientation, and higher positive affect. The pattern of strategic self-regulation, control, and motivation at the high end of the dimension appears to depict a student who fits the description of the good strategy user (Pressley et al., 1987) or the strategic learner (Weinstein et al., 2000).

At the low end of the dimension, the structure coefficients portrayed a pattern of lower self-regulated strategy use, knowledge building, high- and low-level question asking, study time, and study effort. This pattern of lower active strategic self-regulation was associated with a pattern of structure coefficients portraying low control (lower self-efficacy; lower attribution of causality to strategy use/effort, ability [for grades], and help; and lower outcome expectancy for both grades and learning), higher work avoidance and lower mastery and performance approach goal orientations, and lower positive affect. We might have interpreted this dimension as simply unipolar, reflecting increasing levels of motivated strategic self-regulation, except for the strong contribution of work avoidance

Table 2  
Summary Statistics for Canonical Correlation Analysis

Correlation	$R_c$	$R_c^2$	Wilks's $\Lambda$	$\chi^2$	$df$	$p$	Proportion of variance	
							Set 1	Set 2
1	.67	.45	.176	661.92	161	<.001	.36	.12
2	.60	.36	.317	436.67	132	<.001	.12	.09
3	.53	.28	.497	265.73	105	<.001	.13	.11
4	.38	.14	.692	140.34	80	<.001	.10	.04
5	.32	.10	.807	81.71	57	.018	.10	.05
6	.26	.07	.901	39.57	36	.314	.11	.04
7	.18	.03	.967	12.63	17	.761	.07	.05



Table 3  
*Standardized Canonical Coefficients and Structure Coefficients for the Canonical Variates*

Variable	Dimension 1		Dimension 2		Dimension 3	
	Canonical	Structure	Canonical	Structure	Canonical	Structure
Strategic self-regulatory variables						
Self-regulated strategy use	.54	.86	-.53	-.30	-.43	-.13
Knowledge building	.24	.65	.93	.64	.37	.30
High-level question asking	.26	.62	.32	.24	-.63	.13
Low-level question asking	.01	.62	-.39	.09	.74	.28
Lack of regulation	-.00	.14	-.36	-.35	.83	.84
Study time	.08	.53	-.24	-.38	-.01	-.15
Perceived study effort	.30	.56	.12	-.08	-.09	-.08
Control and motivational variables						
Control beliefs						
Self-efficacy	.36	.41	.04	.38	-.08	-.39
Grade success expectancy	-.20	-.02	.19	.49	-.11	-.34
Learning causal attributions						
Strategy/effort	.22	.48	-.12	-.33	-.21	-.11
Ability	.03	.23	-.05	-.10	-.02	.15
Interest/enjoyment	-.20	.09	.10	.25	.14	.23
External causes	-.02	.04	.07	-.13	.25	.52
Help	.17	.44	-.16	-.28	-.02	.32
Grade causal attributions						
Strategy/effort	.00	.47	-.27	-.30	-.15	.00
Ability	.08	.31	-.03	-.14	-.08	.13
Interest/enjoyment	.04	.18	.25	.27	-.06	.17
External causes	-.08	-.01	-.12	-.08	.04	.46
Help	.05	.40	-.07	-.25	.21	.32
Outcome expectancy						
Learning	.05	.37	.49	.47	.32	.49
Grades	.05	.37	-.12	.24	-.01	.41
Locus of control	.02	.13	-.13	-.16	.12	-.13
Goal orientation						
Mastery	.20	.58	.31	.51	-.08	-.03
Performance approach	.11	.43	-.04	.07	.04	.24
Work avoidance	-.39	-.61	.18	.17	.29	.41
Future time perspective						
Connectedness	-.08	.03	.12	-.01	-.10	-.18
Valence	.09	.19	-.07	-.08	.10	.22
Affect						
Positive class affect	.27	.46	.24	.57	.12	.03
Negative class affect	.17	.11	-.07	-.37	.27	.48
Course anxiety	.11	.29	-.15	-.20	.38	.66

Note. Canonical and structure are the standardized canonical coefficients and structure coefficients, respectively.

goal orientation to the low end of the dimension. The higher work avoidance goal orientation suggests a deliberate self-regulatory choice to be strategically inactive and disengaged from the class. Also, this pattern of control, motivation, and strategic self-regulation has been identified in other research, such as that conducted by Reeve et al. (2004), who described the amotivational student as passively "going through the motions" (p. 38), and Moreno, Shell, and Pirritano (2007), who described the apathetic student. We therefore believe that this dimension is bipolar.

*Structure of the second canonical dimension.* At the high end of the second canonical dimension, the structure coefficients portrayed a pattern of higher knowledge-building strategies and lower lack of regulation, but also lower self-regulated strategy use and study time. Canonical coefficients also suggested a contribution from higher high-level and lower low-level question asking. These likely did not have high structure coeffi-

cients because they contributed in opposite directions to the variate, even though they are themselves highly correlated (see Table 1). This pattern of strategic self-regulation was associated with a pattern of structure coefficients portraying higher positive control beliefs (higher self-efficacy and expectancy for success, higher outcome expectancies for learning but not for grades, and lower causal attribution to strategy use/effort), higher mastery goal orientation, and higher positive and lower negative class affect. Canonical coefficients also suggested a contribution to control from higher causal attribution to interest and enjoyment, especially for grades. The high end of this dimension seems to fit the descriptions of the knowledge-building student (Bereiter & Scardamalia, 1989; Chan et al., 1997), the intrinsically motivated, autonomous student described in self-determination theory (Reeve et al., 2004; Ryan & Deci, 2000), or the mastery-oriented student in achievement goal theories (e.g., Pintrich, 2003).

What differentiates these students from those portrayed by Dimension 1 is the lack of general self-regulated strategy use. As these students' knowledge building increased, their self-regulated strategy use and study time decreased. Motivationally, causal attributions to strategic learning/effort, which were strongly associated with Dimension 1, also decreased. Also, outcome expectancy for grades and performance approach goal orientation were absent from Dimension 2. These students appear relatively unconcerned with traditional indicators of success, such as grades or performing well compared to others, even though they expect to get high grades. Also, they do not engage much in the general metacognitive and learning strategies that might be expected to produce higher grades. Rather, they appear to pursue deeper learning for their own growth and development.

At the low end of the dimension, the structure coefficients portrayed a pattern of higher self-regulated strategy use and study time but lower knowledge building and higher lack of regulation. Canonical coefficients also suggested a contribution from higher low-level and lower high-level question asking. This pattern of strategic self-regulation was associated with a pattern of structure coefficients portraying a mixed pattern of control (lower self-efficacy and expectancy for grade success, higher causal attribution to strategic learning/effort, and lower outcome expectancy for learning), lower mastery goal orientation, and lower positive and higher negative class affect. Canonical coefficients suggested a contribution to control from lower causal attribution to interest and enjoyment for both learning and grades.

The second canonical dimension appears to be bipolar, as the low end of the dimension depicts students who fit descriptions of surface learners found in the literature (see Pressley et al., 1987; Weinstein & Mayer, 1986). These students seem to have little personal investment in the class as reflected in the low outcome expectancies for learning and low mastery goal orientation. The students understand that good self-regulated strategy use will lead to success in the course, and they are willing to put in the study time and apply general planning, monitoring, and learning strategies to be successful. However, the students express higher lack of regulation, indicating uncertainty about the effectiveness of their strategy use, likely as a result of their lack of self-efficacy in their ability to effectively apply the strategies they are using and low expectation for high grades. This produces a very rote or surface approach to learning.

*Structure of the third canonical dimension.* At the high end of the third canonical dimension, the structure coefficients portrayed a pattern primarily consisting of higher lack of regulation with some higher knowledge building. The canonical coefficients suggested a contribution from higher low-level and lower high-level question asking as well as lower self-regulated strategy use. This pattern of strategic self-regulation was associated with a pattern of structure coefficients portraying a dysfunctional pattern of control (lower self-efficacy for self-regulation, lower expectancy for grade success, higher outcome expectancy for both learning and grades, and higher causal attribution for both learning and grades to external causes and help), higher work avoidance goal orientation, and higher negative affect and anxiety.

The students portrayed by the high end of this dimension fit descriptions of learned helplessness (Dweck, 1999; Dweck & Leggett, 1988; Stipek & Kowalski, 1989; Weiner, 1986). These students predominantly express lack of regulation, indicating con-

fusion and uncertainty about how to effectively approach studying for the class with corresponding low levels of self-regulated strategy use. When they do ask questions, these students ask low-level questions focused on finding out what the teacher wants and getting rote answers for tests rather than personal meaning-making questions. Their control beliefs are equally dysfunctional. Such students see succeeding in the course and learning the material as important but lack efficacy for executing the strategies and self-regulation need to be successful and hence have little expectation of success. The students attribute causality to external factors, reflecting a lack of personal control over their own success and learning. This lack of control is associated with high negative affect and anxiety. It is probably not surprising that they have only a work avoidance goal orientation.

At the low end of the dimension, the structure coefficients portrayed a pattern of lower lack of regulation and lower knowledge building, with the canonical coefficients showing higher self-regulated strategy use, higher high-level question asking, and lower low-level question asking. This pattern of strategic self-regulation was associated with a pattern of structure coefficients portraying a mixed pattern of control (higher self-efficacy for self-regulation and expectancy for grade success, lower outcome expectancy for both learning and grades, and lower causal attribution for both learning and grades to external causes and help), lower work avoidance goal orientation, and lower levels of negative affect and anxiety.

This pattern of strategic self-regulation, control, and motivation was not theoretically interpretable. This end of the dimension contained aspects of both the strategic learner and the surface learner, but the pattern found did not conform to any strategic or motivational patterns in the existing literature. Therefore, we believe that this dimension is unipolar and reflects only increasing levels of learned helplessness rather than two qualitatively different patterns of self-regulation.

### *Research Question 1: Are There Different Dimensions of Control Associated With Different Patterns of Strategic Self-Regulation?*

Although we found strong support for multiple control and strategic self-regulation dimensions, our hypothesis that competency and contingency beliefs would form distinct control dimensions was not supported. The canonical correlation analysis found three distinct dimensions that contained five patterns of control. Two of these five patterns reflected positive perceived control. The good strategy user end of Dimension 1 conformed to classic formulations of high perceived control (e.g., Chapman et al., 1990; Ford & Thompson, 1985; Heckhausen & Schulz, 1995; Skinner, 1996; Skinner et al., 1988; Walls & Little, 2005; Weisz & Stipek, 1982). This pattern consisted of high self-efficacy for learning strategies, causal attribution to strategy use/effort and help for both grades and learning and to ability for grades, and high outcome expectancy for learning and grades. The only deviation from the expected control pattern was a lack of expectancy for grade success. This pattern of perceived control was associated, as would be expected, with a strategic self-regulatory pattern of high self-regulated strategy use, knowledge building, question asking, study time, and study effort.



The knowledge-building end of Dimension 2, however, identified a second positive perceived control pattern. This pattern consisted of high self-efficacy for learning strategies, high expectancy for grade success, and high outcome expectancy for learning but not grades. There was high causal attribution to internal causes, but to interest and enjoyment rather than to strategy use/effort or ability. In fact, attribution to strategy use/effort contributed negatively. This pattern of control was linked to a more specific strategic self-regulatory pattern of higher knowledge-building strategies and high-level question asking.

Three control patterns were more negative. The pattern of control in the learned helplessness end of Dimension 3 has been noted previously as being particularly dysfunctional and psychologically damaging (e.g., Bandura, 1997; Dweck & Leggett, 1988; Skinner, 1996; Stipek & Kowalski, 1989; Weiner, 1986). This pattern consisted of high outcome expectancy for learning and grades but low self-efficacy and expectancy for grade success along with causal attributions to external causes and help. Students with this pattern of control see the class as important but see themselves as incapable of effectively exerting control and learning. This pattern of control was associated with a strategic self-regulatory pattern of high lack of regulation and low-level question asking and low self-regulated strategy use and high-level question asking. In something of an anomaly, this pattern also contained higher knowledge-building strategies. This may reflect an association between these strategies and the high learning outcome expectancy in the control pattern.

A second pattern of low personal control was evidenced in the apathetic end of Dimension 1. This pattern consisted of low self-efficacy; low outcome expectancies for both learning and grades; and low causal attributions to strategy use/effort, help, and ability. As would be expected in control theory (Skinner, 1996), this almost complete absence of any sense of personal control was associated with a strategic self-regulatory profile of virtually no strategy use, knowledge building, question asking, or studying. Unlike the learned helpless pattern, however, this self-regulatory pattern had no contribution from lack of regulation, suggesting that not doing these self-regulatory behaviors was not due to a lack of understanding what to do.

The final pattern of somewhat mixed control was found in the surface learning end of Dimension 2. This pattern consisted of low self-efficacy, expectancy for grade success, and outcome expectancy for both learning and grades, like the apathetic pattern, but contained high causal attributions to strategy use/effort and help, like the good strategy user pattern. This control pattern was associated with a strategic self-regulatory pattern of higher self-regulated strategy use, low-level question asking, and study time, but also higher lack of regulation and lower knowledge-building strategies and high-level question asking. This control pattern suggests a belief that classroom success depends on one's own strategy use and effort but a lack self-efficacy in one's ability to do these strategies effectively. This is then manifest in higher self-regulated learning strategies and study time, but also in higher lack of regulation, suggesting an uncertainty about how to effectively apply the strategies being attempted.

Although control beliefs did not form unique competency and contingency dimensions, patterns of control consistent with the competency and contingency framework (Shell & Husman, 2001; Weisz & Stipek, 1982) were apparent. Regardless of dimension or

pattern of contingency beliefs, high self-efficacy for learning strategies was always part of a control pattern associated with a strategic self-regulatory pattern that would be expected to produce higher achievement. More specifically, as hypothesized by Shell and Husman (2001), higher self-efficacy was always part of a control pattern associated with a strategic self-regulatory pattern containing higher knowledge-building strategies and high-level question asking. Lower self-efficacy was always part of a control pattern associated with a self-regulatory pattern that would be expected to result in lower achievement. Specifically, lower self-efficacy for learning strategies was always part of a control pattern associated with higher lack of regulation.

Conversely, contingency beliefs were part of control patterns associated with all strategic self-regulatory patterns except the apathetic end of Dimension 1; so, as found by Shell and Husman (2001), they were not differentially associated with self-regulatory patterns likely to produce higher and lower achievement. However, the association of contingency beliefs with reported study time, study effort, and self-regulated strategy use mirrored the associations found by Shell and Husman. In all dimensions, higher attribution to strategy use/effort and higher outcome expectancy were part of a control pattern associated with a strategic self-regulatory pattern containing higher studying and higher use of effortful strategic self-regulatory strategies. These findings strengthen the hypothesis that contingency beliefs motivate effort and general level of strategic behavior, whereas competency or self-efficacy beliefs more specifically motivate the use of more cognitively complex deep processing strategies.

### *Research Question 2: Are Different Dimensions of Control Connected to Different Goal Orientations?*

We found partial support for our hypothesis that control beliefs would form unique dimensions differentiated by mastery, performance approach, and work avoidance goal orientations. Two of the three identified dimensions had contributions from only a single goal orientation. Mastery goal orientation was the only goal orientation contributing to Dimension 2 (knowledge building/surface learning), and work avoidance goal orientation was the only goal orientation contributing to Dimension 3 (learned helplessness). Dimension 1 (good strategy user/apathetic) had contributions from all three goal orientations: mastery, performance approach, and work avoidance.

Regardless of the canonical dimension, mastery goal orientation always contributed to a dimension in conjunction with a pattern of high perceived control, and work avoidance goal orientation always contributed to a dimension in conjunction with a pattern of low or dysfunctional perceived control. Performance approach goal orientation contributed to a dimension in conjunction with a pattern of high perceived control in the good strategy user end of Dimension 1 but did not contribute to the other dimensions. These connections between goal orientation and perceived control generally correspond to what would be expected on the basis of theory and previous research (e.g., Martin et al., 2003; Midgley et al., 2001; Pekrun et al., 2006; Pintrich, 2000a, 2000b; Skinner, 1996; Wolters & Yu, 1996)



*Research Question 3: How Do Control Beliefs for Learning Versus Achievement Outcomes Differ in Their Influence on Strategic Self-Regulation, and Are These Differences Connected to Mastery and Performance Approach Goal Orientations?*

We found few differences in the contributions to the canonical variates for control beliefs for learning and achievement outcomes and no support for our hypothesis that control beliefs for learning and grades would be more strongly connected to mastery and performance approach goal orientations, respectively. The contributions from students' causal attributions for learning and grades paralleled each other in all dimensions. Although there were no differential contributions for learning and grade attributions, causal attributions did contribute differentially to the canonical dimensions in conjunction with different goal orientations. Attributions to effort and strategy use contributed to Dimension 1 (good strategy user/apathetic) in the same direction as the contributions from both mastery and performance approach goal orientations but contributed in the opposite direction to mastery goal orientation in Dimension 2 (knowledge building/surface learning). Apparently, students at the knowledge-building end of Dimension 2 do not see pursuit of mastery-oriented goals and personal knowledge building as being caused by effort and learning strategies, whereas students at the surface learning end of this dimension attribute causality to effort and strategy use and do not pursue mastery-oriented goals. Attribution to help contributed to Dimension 1 (good strategy user/apathetic) in the same direction as mastery and performance approach goal orientations and in the opposite direction to work avoidance goal orientation but contributed to Dimension 3 (learned helplessness) in the same direction as work avoidance goal orientation. External causes, as expected, contributed to Dimension 3 (learned helplessness) in the same direction as work avoidance goal orientation.

Unlike for causal attributions, there were differences in how students' learning and grade outcome expectancies contributed to the dimensions. In Dimension 1 (good strategy user/apathetic) and Dimension 3 (learned helplessness), the contributions from learning and grade outcome expectancies paralleled each other. However, in these dimensions, their contributions did not parallel the contributions from any specific goal orientation, nor were they associated consistently with aspects of strategic self-regulation. In Dimension 1 (good strategy user/apathetic), higher outcome expectancies contributed in the same direction as mastery and performance goal approach goal orientations and in the opposite direction to work avoidance goal orientation and were associated with higher effort and self-regulated strategy use, but in Dimension 3 (learned helplessness), they contributed in the same direction as work avoidance goal orientation and were associated with lower effort and general strategy use. In Dimension 2 (knowledge building/surface learning), only higher learning outcome expectancies contributed to the dimension. This contribution was in the same direction as mastery goal orientation and was associated with knowledge building but not with effort and general strategy use. Overall, outcome expectancies did not appear to be uniquely linked with any specific goal orientation.

Self-efficacy for learning strategies always contributed to a dimension in the same direction as mastery goal orientation, suggesting a strong link between efficacy for learning and being

oriented toward setting learning or mastery goals. In Dimension 1 (good strategy user/apathetic), self-efficacy for learning strategies contributed in the same direction as performance approach goal orientation, but self-efficacy for learning and performance approach goal orientation did not contribute similarly to other dimensions, suggesting a less strong link with self-efficacy than for mastery goal orientation. In Dimension 1 (good strategy user/apathetic) and Dimension 3 (learned helplessness), self-efficacy for learning strategies contributed in the opposite direction as work avoidance goal orientation. Unexpectedly, success expectancy for grades, which reflects belief about an achievement outcome, did not contribute to any dimension with performance approach goal orientation. In Dimension 2 (knowledge building/surface learning), it contributed in the same direction as higher mastery goal orientation, and in Dimension 3 (learned helplessness), it contributed in the same opposite direction as work avoidance goal orientation.

*Research Question 4: How Are Dimensions of Control Beliefs Connected to FTP?*

We found no support for our hypothesis that higher perceived control would be connected to higher FTP. Unlike previous findings (Shell & Husman, 2001), neither FTP valence nor FTP connectedness contributed meaningfully to any of the three canonical dimensions. A key difference between the Shell and Husman (2001) study and the present study, however, is that this study was contextualized within a single class, whereas Shell and Husman examined overall college GPA and general study time and effort for all classes. Theoretically, FTP is a dispositional psychological construct that is decontextual. Thus, the FTP variables examined may not have been as strongly associated with the class-specific strategic, self-regulatory behaviors as the class-specific control, goal orientation, and affective variables were, even if these FTP variables are associated with academic achievement and strategic self-regulation in general. These results suggest that further research is needed to clarify how variables at different levels of decontextualization interact in classroom settings.

*Research Question 5: How Are Dimensions of Control Connected to Different Patterns of Affect?*

Consistent with Pekrun et al. (2006), we found support for our hypothesis that higher perceived control would be connected to more positive affect. A pattern of high perceived control contributed in the same direction as positive class affect in Dimension 1 (good strategy user/apathetic) and Dimension 2 (knowledge building/surface learning). A pattern of low or dysfunctional perceived control contributed in the same direction as more negative affect in Dimension 3 (learned helplessness) and Dimension 2 (knowledge building/surface learning).

We also found support for Pekrun et al.'s (2006) findings that mastery and performance approach goal orientation are linked to different control perceptions and subsequent emotions. Mastery goal orientation always contributed to a dimension in conjunction with higher perceived control and positive class affect. Like Pekrun et al., we found that the performance approach goal orientation in this study also contributed in conjunction with higher perceived control and positive affect in Dimension 1 (good strat-



egy user/apathetic). Work avoidance goal orientation acted similarly to performance avoidance goal orientation in Pekrun et al.'s study. Higher work avoidance goal orientation contributed in the same direction as the dysfunctional control pattern and negative affect in Dimension 3 (learned helplessness) and the low perceived control and low positive affect in the apathetic end of Dimension 1.

Anxiety was not as clearly associated with perceived control. Consistent with previous control literature (Pekrun et al., 2006; Skinner, 1996), higher anxiety contributed in the same direction as dysfunctional control in Dimension 3 (learned helplessness) but did not contribute to any other dimension. The structure coefficient for anxiety was just below our cutoff in Dimension 1 (good strategy user/apathetic), suggesting, as Zeidner (1998) suggested, a potential positive role for anxiety in motivating control and strategic self-regulation, but this requires further research to substantiate.

## Discussion

### *The Multivariate Dimensionality of Control*

Our goal in this study was to further clarify the multidimensional nature of control in academic classroom settings by conducting a systematic, multidimensional examination of how control beliefs, in conjunction with other goal, motivational, and affect constructs, were related to academic strategic self-regulation. The results show that there is a dimensionality in the relations between control and strategic self-regulation that is distinct from the factorial dimensionality found for the component control beliefs themselves (e.g., Chapman et al., 1990; Skinner, 1996; Skinner et al., 1988; Walls & Little, 2005). Within these relationship dimensions, the association of any specific control belief to strategic self-regulation is dependent on the overall pattern of the other control beliefs, goal orientations, and affect present in the dimension.

Dimension 1 fit the traditional view of control (Chapman et al., 1990; Skinner, 1996; Skinner et al., 1988; Walls & Little, 2005; Weisz & Stipek, 1982). High control consisting of high agency (self-efficacy), high response–outcome contingencies (outcome expectancies), and internal means–ends contingencies (attribution to effort and ability) was associated with high use of self-regulated learning strategies, knowledge building, question asking, study time, and perceived study effort. Low control consisting of the absence of these beliefs was associated with no self-regulated strategy use, study effort, or active classroom engagement through question asking.

Dimension 3 also identified a well-established dysfunctional control pattern of low agency and agent–ends contingencies with means–ends contingencies for external causes paired with high response–outcome contingencies (outcome expectancies). Theorists (Bandura, 1997; Dweck & Leggett, 1988; Skinner, 1996; Stipek & Kowalski, 1989; Weiner, 1986) have noted the problematic nature of this pattern of control beliefs, which produces a learned-helplessness approach to self-regulation.

Dimension 2, however, identified patterns of control that have not been extensively discussed in the literature. Control in Dimension 2 seemed to contrast with effort. In general, control is thought to increase effort (Schmitz & Skinner, 1993; Shell & Husman, 2001; Skinner, 1996). In Dimension 2, however, higher agency,

agent–ends contingencies, and response outcome expectancies for learning were associated with more deep knowledge-building strategies without corresponding increases in general study time and effort or use of more general self-regulated learning strategies. Conversely, at the low end of the dimension, high means–ends contingencies to strategy use and effort, but low agency, agent–ends contingencies, and response outcome expectancies for learning were associated with exerting effort through higher use of general self-regulated learning strategies, study time, and low-level question asking but lower knowledge building and high-level question asking.

In relation to patterns of control, positive patterns of control appear to always include high self-efficacy for learning strategies and high outcome expectancy for learning. No other control belief was present in all positive control patterns. More negative or dysfunctional patterns of control appear always to have low self-efficacy for learning strategies, but no other control belief was present or absent in all negative patterns. Only three generalizations about relations between control and strategic self-regulation appeared to hold. First, higher self-efficacy for learning (agency) was always present in control patterns associated with more personal knowledge building and high-level question asking. Second, higher causal attribution to strategy use and effort was always present in control patterns associated with more use of self-regulated learning strategies and more study time. Interestingly, this held even if students had low self-efficacy for effectively executing the strategies they were using, as in the surface learning end of Dimension 2. Third, higher outcome expectancies were always present in control patterns associated with more positive strategic self-regulatory patterns, but not exclusively, as they also were in the control pattern associated with learned helplessness.

The gist of these findings is that the answer to “How does control influence academic strategic self-regulation?” is not straightforward. The answer depends on what pattern of control is involved and also on what type of strategic self-regulation one is referring to, especially effort expenditure versus deeper strategy use. The finding that effort expenditure is not always associated with every pattern of positive control perceptions and that control-motivated effort and strategy use may result in only surface level learning may help explain some of the anomalies in the link between control, effort expenditure, and achievement found by Schmitz and Skinner (1993). Like their results, the findings suggest that higher control does not necessarily lead to more effort, and neither does more effort lead to a strategic self-regulated pattern that is optimal.

### *The Interaction of Control, Goals, FTP, and Affect*

The results help fill the gap in empirical studies of the linkages between control and other motivational constructs noted by Walls and Little (2005). Control appears to be intertwined with both students' goal orientation and affect. Perception of high control always occurred in conjunction with mastery goal orientation and positive affect. These findings strengthen arguments for the primacy of mastery-oriented goals in achievement goal theory (e.g., Martin et al., 2003; Midgley et al., 2001; Pintrich, 2000a, 2000b, 2003). Conversely, work avoidance goal orientation always occurred in conjunction with low control perceptions, especially low self-efficacy, and in the case of learned helplessness, with a



dysfunctional pattern of high outcome expectancy and low self-efficacy. As hypothesized in self-determination theory (Reeve et al., 2004; Ryan & Deci, 2000), students who lack control do not appear to be motivated to set productive goals.

Performance approach goals were not strongly connected to patterns of control. The performance approach goals in this study occurred in conjunction with higher control only in Dimension 1 (good strategy user/apathetic). These findings support contentions that performance approach goals can be positive influences on students' self-regulation (Midgley et al., 2001; Pintrich, 2000a) but suggest that positive benefits are tied also to having mastery goals and to having high perceptions of perceived control. We did not assess performance avoidance goals, but the results suggest a potential role for these goals. The surface-learning end of Dimension 2 had a negative contribution from mastery goal orientation but did not have a positive contribution from any goal orientation. Looking at the rather negative control pattern of low self-efficacy and expectancy for success along with low outcome expectancies for learning or grades that characterized the surface-learning end of the dimension, we can speculate that the identified pattern of high causal attributions for strategy use and effort coupled with high self-regulated strategy use and study time could be supported by a performance avoidance goal of not appearing incompetent or stupid. These students did not appear to be studying and exerting effort for a positive reason, so perhaps they were doing so for a negative reason. This clearly deserves further study.

The findings strongly supported the connections between goal orientation, control, and affect proposed by Pekrun in his control-value theory of achievement emotion (Pekrun et al., 2006, 2002), as well as similar theoretical perspectives in self-determination theory (Reeve et al., 2004; Ryan & Deci, 2000) and self-worth theory (Covington, 2004). Positive affect appears to occur in conjunction with control exercised in the pursuit of mastery-oriented goals. No other combination of control and goal orientation was linked with high positive affect. Performance approach goal orientation did not appear to be uniquely linked with affect, as it only had a meaningful contribution to a dimension in conjunction with mastery goal orientation. Work avoidance goal orientation contributed in conjunction with high negative affect and anxiety in Dimension 3 (learned helplessness) and with lower positive affect in the apathetic end of Dimension 1. Work avoidance goal orientation also may have contributed in conjunction with lower anxiety (in opposition to its contribution with anxiety in Dimension 3) in the apathetic end of Dimension 1, but this potential linkage is only speculative because of the only marginal structure coefficient for anxiety in this dimension.

### *Dispositional Versus Class-Specific Beliefs*

Contrary to previous findings (De Volder & Lens, 1982; Miller et al., 1996; Husman et al., 2004; Shell & Husman, 2001), our findings were that FTP did not contribute meaningfully to any of the three canonical dimensions. Neither did locus of control, which has been shown to be associated with academic achievement in many studies (e.g., Heckhausen & Schulz, 1995; Shell & Husman, 2001; Stipek & Weisz, 1981). As noted by theorists (Bandura, 2001; Stipek & Weisz, 1981), a possible explanation for this is that control and other motivational constructs are psychologically hierarchical. From this perspective, students' self-regulated strategic

learning in a class would be most proximally motivated by their class-specific control perceptions, goal orientations, and affect, and these proximal motivators might wash out any influence from more dispositional, general beliefs like FTP, locus of control, or other self-system constructs. However, the proximal motivators themselves might be influenced by these dispositional-level constructs. These hierarchical patterns of influences deserve further study.

### *Final Conclusions*

The results suggest that motivation for academic strategic self-regulation derived from control, goal orientation, and affect is complex. Increasing or decreasing specific control beliefs or even the overall level of perceived control does not lead directly to any specific level of students' effort, strategy use, or self-regulation. Also, the influence of control is embedded within students' goal orientation and affective reactions to a course. We did not identify patterns of students' control-based motivation or strategic self-regulation that have not appeared in the literature; however, the findings help clarify how specific self-regulatory patterns, including negative or dysfunctional patterns, are motivated by specific constellations of control, goal orientation, and affect. The associations between these constellations and student strategic self-regulation do not fall along a single continuum. Rather, constellations of control, goal orientation, and affect appear to have qualitatively distinct multiple continuums of association, in which specific control beliefs, goal orientations, or affect may have different relationships to specific strategic self-regulatory behaviors and effort in one continuum than in another.

This study is limited in that we only examined college students. Future research is needed to determine whether similar dimensional continuums and relations between control and strategic self-regulation are present for younger students. The results confirm the need for the systematic, multivariate examination of control and other motivational constructs called for by control theorists (Little, 1998; Skinner, 1996; Skinner et al., 1988; Walls & Little, 2005). We could not have identified the canonical dimensions without a comprehensive inclusion of control, goal orientation, and affect measures, as well as a broad sampling of strategic self-regulatory and effort indicators. This multivariate examination needs to be continued and expanded to include other important motivational constructs that have been identified in the literature (e.g., Pintrich, 2003).

### *References*

- Ames, C. A. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*, 261–271.
- Atkinson, J. (1964). *An introduction to motivation*. Princeton, NJ: Van Nostrand.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.
- Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology, 52*, 1–26.
- Bereiter, C., & Scardamalia, M. (1989). Intentional learning as a goal of instruction. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 361–392). Hillsdale, NJ: Erlbaum.
- Biggs, J., Kember, D., & Leung, D. Y. P. (2001). The revised two-factor



- study process questionnaire: R-SPQ-2F. *British Journal of Educational Psychology*, 71, 133–149.
- Boekaerts, M., Pintrich, P. R., & Zeidner, M. (Eds.). (2000). *The handbook of self-regulation*. San Diego, CA: Academic Press.
- Brown, A. L., Bransford, J., Ferrara, R., & Campione, J. (1983). Learning, remembering, and understanding. In P. H. Musen (Ed.), *Handbook of child psychology* (pp. 77–166). New York: Wiley.
- Chan, C., Burtis, J., & Bereiter, C. (1997). Knowledge building as a mediator of conflict in conceptual change. *Cognition and Instruction*, 15, 1–40.
- Chapman, M., Skinner, E. A., & Baltes, P. B. (1990). Interpreting correlations between children's perceived control and cognitive performance: Control, agency, or means–ends beliefs. *Developmental Psychology*, 26, 246–253.
- Covington, M. V. (2004). Self-worth theory: Goes to college or do our motivation theories motivate? In D. M. McInerney & S. Van Etten (Eds.), *Big theories revisited: Research on sociocultural influences on motivation and learning* (Vol. 4, pp. 91–114). Greenwich, CT: Information Age Publishing.
- De Volder, M. L., & Lens, W. (1982). Academic achievement and future time perspective as a cognitive-motivational concept. *Journal of Personality and Social Psychology*, 42, 566–571.
- Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. New York: Psychology Press.
- Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, 95, 256–273.
- Ford, M. E., & Thompson, R. A. (1985). Perceptions of personal agency and infant attachment: Toward a life-span perspective on competence development. *International Journal of Behavioral Development*, 8, 377–406.
- Hamilton, R. J., & Akhter, S. (2002). Psychometric properties of the multidimensional-multiattributational causality scales. *Educational and Psychological Measurement*, 62, 802–817.
- Hamman, D. D. (1998). Preservice teachers' value for learning-strategy instruction. *Journal of Experimental Education*, 66, 209–221.
- Heckhausen, J., & Schulz, R. (1995). A life-span theory of control. *Psychological Review*, 102, 284–304.
- Husman, J., Derryberry, W. P., Crowson, H. M., & Lomax, R. (2004). Instrumentality, task value, and intrinsic motivation: Making sense of their independent interdependence. *Contemporary Educational Psychology*, 29, 63–76.
- Husman, J., & Lens, W. (1999). The role of the future in student motivation. *Educational Psychologist*, 34(2), 113–125.
- Lens, W. (1987). Future time perspective, motivation and school performance. In E. De Corte, H. Lodewijks, R. Parmentier, & P. Span (Eds.), *Learning and instruction: European research in an international context* (Vol. 1, pp. 181–189). Elmsford, NY: Pergamon Press.
- Little, T. D. (1998). Sociocultural influences on the development of children's action-control beliefs. In J. Heckhausen & C. S. Dweck (Eds.), *Action and self-development: Theory and research through the life span* (pp. 3–36). Thousand Oaks, CA: Sage.
- Martin, A. J., Marsh, H. W., & Debus, R. L. (2003). Self-handicapping and defensive pessimism: A model of self-protection from a longitudinal perspective. *Contemporary Educational Psychology*, 28, 1–36.
- Midgley, C., Kaplan, A., & Middleton, M. (2001). Performance-approach goals: Good for what, for whom, under what circumstances, and at what cost? *Journal of Educational Psychology*, 93, 77–86.
- Miller, R. B., Greene, B. A., Montalvo, G. P., Ravindran, B., & Nichols, J. D. (1996). Engagement in academic work: The role of learning goals, future consequences, pleasing others, and perceived ability. *Contemporary Educational Psychology*, 21, 388–422.
- Moreno, R., Shell, D. F., & Pirritano, M. (2007, April). *Factors predictive of mathematics performance: Diversity between and within White, Hispanic, and Native American children*. Paper presented at the American Educational Research Association Annual Meeting, Chicago, IL.
- Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, 66, 543–578.
- Pekrun, R., Elliot, A. J., & Maier, M. A. (2006). Achievement goals and discrete achievement emotions: A theoretical model and prospective test. *Journal of Educational Psychology*, 98, 583–597.
- Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, 37, 91–105.
- Pintrich, P. R. (2000a). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology*, 92(3), 544–555.
- Pintrich, P. R. (2000b). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *The handbook of self-regulation* (pp. 451–502). San Diego, CA: Academic Press.
- Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology*, 95(3), 667–686.
- Pintrich, P. R. (2004). A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review*, 16, 385–407.
- Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research and practice*. Englewood Cliffs, NJ: Merrill.
- Pintrich, P. R., Smith, D., Garcia, T., & McKeachie, W. (1993). Predictive validity and reliability of the Motivated Strategies for Learning Questionnaire (MSLQ). *Educational and Psychological Measurement*, 53, 801–813.
- Pressley, M., Borkowski, J. G., & Schneider, W. (1987). Cognitive strategies: Good strategy users coordinate metacognition and knowledge. In R. Vasta & G. Whitehurst (Eds.), *Annals of child development* (Vol. 5, pp. 89–129). Greenwich, CT: JAI Press.
- Reeve, J., Deci, E. L., & Ryan, M. R. (2004). Self-determination theory: A dialectical framework for understanding sociocultural influences on student motivation. In D. M. McInerney & S. Van Etten (Eds.), *Big theories revisited: Research on sociocultural influences on motivation and learning* (Vol. 4, pp. 31–60). Greenwich, CT: Information Age Publishing.
- Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychology Monographs*, 80, 1–28.
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55, 68–78.
- Scardamalia, M., & Bereiter, C. (1992). Text-based and knowledge-based questioning by children. *Cognition and Instruction*, 9, 177–199.
- Schmitz, B., & Skinner, E. (1993). Perceived control, effort, and academic performance: Interindividual, intraindividual, and multivariate time-series analysis. *Journal of Personality and Social Psychology*, 64, 1010–1028.
- Schraw, G., Horn, C., Thorndike-Christ, T., & Bruning, R. (1995). Academic goal orientations and student classroom achievement. *Contemporary Educational Psychology*, 20, 359–368.
- Schunk, D. H., & Pajares, F. (2004). Self-efficacy in education revisited: Empirical and applied evidence. In D. M. McInerney & S. Van Etten (Eds.), *Big theories revisited: Research on sociocultural influences on motivation and learning* (Vol. 4, pp. 115–138). Greenwich, CT: Information Age Publishing.
- Shell, D. F. (2005). [Analysis of the reliability of SPOCK scales]. Unpublished raw data.
- Shell, D. F., Colvin, C., & Bruning, R. H. (1995). Self-efficacy, attribution, and outcome expectancy mechanisms in reading and writing achievement: Grade level and achievement level differences. *Journal of Educational Psychology*, 87, 386–398.

- Shell, D. F., & Husman, J. (2001). The multivariate dimensionality of personal control and future time perspective in achievement and studying. *Contemporary Educational Psychology*, 26, 481–506.
- Shell, D. F., Husman, J., Turner, J. E., Cliffler, D. M., Nath, I., & Sweany, N. (2005). The impact of computer-supported collaborative learning communities on high school students' knowledge building, strategic learning, and perceptions of the classroom. *Journal of Educational Computing Research*, 33(3), 327–349.
- Shell, D. F., Murphy, C. C., & Bruning, R. H. (1989). Self-efficacy and outcome expectancy mechanisms in reading and writing achievement. *Journal of Educational Psychology*, 81(1), 91–100.
- Skinner, E. A. (1996). A guide to constructs of control. *Journal of Personality and Social Psychology*, 71, 549–570.
- Skinner, E. A., Chapman, M., & Baltes, P. B. (1988). Control, means-ends, and agency beliefs: A new conceptualization and its measurement during childhood. *Journal of Personality and Social Psychology*, 54, 117–133.
- Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Erlbaum.
- Stipek, D. J., & Kowalski, P. S. (1989). Learned helplessness in task-orienting versus performance-orienting testing conditions. *Journal of Educational Psychology*, 81(3), 384–391.
- Stipek, D. J., & Weisz, J. R. (1981). Perceived personal control and academic achievement. *Review of Educational Research*, 51, 101–137.
- Thompson, B. (1984). *Canonical correlation analysis: Uses and interpretation* (Sage University Paper series on quantitative applications in the social sciences, Series no. 07–047). Beverly Hills, CA: Sage.
- Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education* (6th ed.). Upper Saddle River, NJ: Prentice Hall.
- Vermunt, J. D., & Vermetten, Y. J. (2004). Patterns in student learning: Relationships between learning strategies, conceptions of learning, and learning orientations. *Educational Psychology Review*, 16, 359–384.
- Vispoel, W. P., & Austin, J. R. (1995). Success and failure in junior high school: A critical incident approach to understanding students' attributional beliefs. *American Educational Research Journal*, 32, 377–412.
- Walls, T. A., & Little, T. D. (2005). Relations among personal agency, motivation, and school adjustment in early adolescence. *Journal of Educational Psychology*, 97, 23–31.
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS Scales. *Journal of Personality and Social Psychology*, 54, 1063–1070.
- Weiner, B. (1986). *An attributional theory of motivation and emotion*. New York: Springer-Verlag.
- Weiner, B. (2004). Attribution theory revisited: Transforming cultural plurality into theoretical unity. In D. M. McInerney & S. Van Etten (Eds.), *Big theories revisited: Research on sociocultural influences on motivation and learning* (Vol. 4, pp. 13–30). Greenwich, CT: Information Age Publishing.
- Weinstein, C. E., Husman, J., & Dierking, D. R. (2000). Interventions with a focus on learning strategies. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 727–747). San Diego, CA: Academic Press.
- Weinstein, C. E., & Mayer, R. E. (1986). The teaching of learning strategies. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 315–327). New York: Macmillan.
- Weinstein, C., Zimmermann, S., & Palmer, D. (1988). Assessing learning strategies: The design and development of the LASSI. In C. Weinstein, E. Goetz, & P. Alexander (Eds.), *Learning and study strategies: Issues in assessment, instruction, and evaluation* (pp. 279–306). Hillsdale, NJ: Erlbaum.
- Weisz, J. R., & Stipek, D. J. (1982). Competence, contingency, and the development of perceived control. *Human Development*, 25, 250–281.
- Wigfield, A., & Eccles, J. S. (2002). *Development of achievement motivation*. San Diego, CA: Academic Press.
- Wolters, C. A. (2003). Understanding procrastination from a self-regulated learning perspective. *Journal of Educational Psychology*, 95, 179–187.
- Wolters, C. A., & Yu, S. L. (1996). The relation between goal orientation and students' motivational beliefs and self-regulated learning. *Learning & Individual Differences*, 8, 211–238.
- Zeidner, M. (1998). *Test anxiety: The state of the art*. New York: Plenum Press.
- Zimmerman, B. J. (2004). Sociocultural influence and students' development of academic self-regulation: A social-cognitive perspective. In D. M. McInerney & S. Van Etten (Eds.), *Big theories revisited: Research on sociocultural influences on motivation and learning* (Vol. 4, pp. 139–164). Greenwich, CT: Information Age Publishing.
- Zimmerman, B. J., & Bandura, A. (1994). Impact of self-regulatory influences on writing course attainment. *American Educational Research Journal*, 31, 845–862.
- Zimmerman, B. J., & Martinez-Pons, M. (1988). Construct validation of a strategy model of student self-regulated learning. *Journal of Educational Psychology*, 80, 284–290.
- Zimmerman, B. J., & Schunk, D. H. (Eds.). (2001). *Self-regulated learning and academic achievement: Theoretical perspectives* (2nd ed.). Mahwah, NJ: Erlbaum.

Received March 30, 2005

Revision received October 25, 2007

Accepted October 25, 2007 ■



# What Makes Lessons Interesting? The Role of Situational and Individual Factors in Three School Subjects

Yi-Miau Tsai, Mareike Kunter, Oliver Lüdtke, and  
Ulrich Trautwein  
Max Planck Institute for Human Development

Richard M. Ryan  
University of Rochester

The present study investigated intraindividual variation in students' interest experience in 3 school subjects and the predictive power of perceived autonomy support and control. Participants were 261 students in 7th grade. After a survey of students' individual interests and other individual characteristics, repeated lesson-specific measures of students' interest experience and perceived autonomy support and control during instruction were obtained over a 3-week period. Hierarchical linear modeling showed 36%–45% of the variance to be located at the within-student level. Moreover, perceived autonomy support and control during lessons, as well as individual interest, predicted students' interest experience in the classroom.

**Keywords:** interest experience, individual interest, autonomy support, self-determination theory, repeated measurement design

Five minutes before the end of a lesson, students may be waiting impatiently for the bell to ring or be so engaged in the lesson that they are quite unaware of the time. Most students are probably familiar with both experiences. Whether or not a lesson was interesting is one of the key dimensions on which students judge their experience in the classroom. Indeed, the psychological state of being interested plays a major role in students' motivation and learning (Pintrich, 2003a; Urdan & Turner, 2005). Interest has been found to be associated with focused attention, higher cognitive functioning, and learning (Ainley, Hidi, & Berndorff, 2002; Köller, Baumert, & Schnabel, 2003; Krapp, Hidi, & Renninger, 1992). However, a precise understanding of how students' interest experiences emerge in classroom settings is lacking (Pintrich, 2003b). Moreover, although interest experience is influenced by a combination of stable individual characteristics and aspects of the current situation, most research to date has investigated these two sources of influence separately. The purpose of the present study is to investigate both sources of influence directly by assessing students' interest experience in different classroom learning situations. In particular, we examine how individual students' interest experience varies from lesson to lesson within a subject domain. Furthermore, we investigate whether interest experience can be

predicted by stable individual characteristics and by situational aspects that are specific to each lesson.

In the following sections, we first introduce the construct of *interest experience* as a temporary psychological state. We then discuss stable individual characteristics that are assumed to influence interest experience. Turning to situational factors, we next draw on the framework of self-determination theory to outline features of the instructional process that are theorized to enhance interest, focusing on various forms of autonomy support.

## Interest Experience

Interest experience is a psychological state that is characterized by an affective component of positive emotion and a cognitive component of concentration (Hidi & Renninger, 2006). When persons experience interest, their actions acquire an intrinsic quality; they are driven by enjoyment rather than external reasons (Krapp, 2002b). Interest is central to both intrinsic motivation and autonomous forms of extrinsic motivation (Deci, 1992; Ryan & Connell, 1989), and extreme forms of interest experience may be regarded as "flow" (Csikszentmihalyi, 1975).

States of interest arise through an interaction between the person and the surrounding context (Bergin, 1999; Sansone & Thoman, 2005). Consequently, several researchers have proposed that both situational and individual factors should be taken into account when trying to explain different levels of interest (Ainley, Hillman, & Hidi, 2002; Bergin, 1999; Hidi & Renninger, 2006; Sansone & Thoman, 2005). In the framework proposed by Krapp (2002b), the state of interest is assumed to be a function of two classes of distinct influences. The first influence comes from characteristics of the person him- or herself: stable features such as gender, prior knowledge, and experience and relatively stable preferences for certain content areas ("individual interest"). The second influence comes from characteristics of the situation. Certain features of the learning situation are assumed to be capable of arousing the individual's curiosity or interest, regardless of personal prefer-

---

Yi-Miau Tsai, Mareike Kunter, Oliver Lüdtke, and Ulrich Trautwein, Center for Educational Research, Max Planck Institute for Human Development, Berlin, Germany; Richard M. Ryan, Department of Clinical and Social Sciences in Psychology, University of Rochester.

This article is in partial fulfillment of Yi-Miau Tsai's doctoral dissertation requirements within the International Max Planck Research School on the Life Course: Evolutionary and Ontogenetic Dynamics. We thank Jacquelynne Eccles and Jürgen Baumert for their guidance and feedback throughout this research and Susannah Goss for editorial assistance.

Correspondence concerning this article should be addressed to Yi-Miau Tsai, Max Planck Institute for Human Development, Lentzeallee 94, 14195, Berlin, Germany. E-mail: tsai@mpib-berlin.mpg.de

ences (the “interestingness” of a situation). Examples of situational features assumed to trigger interest in the classroom include games and humor (Bergin, 1999).

Several interest researchers have used the term *situational interest* to refer to the psychological state of interest (e.g., Hidi & Anderson, 1992; Schraw & Lehman, 2001). The psychological state described—that is, a state of positive emotion and heightened concentration—is congruent with our description of interest experience. However, research on situational interest has tended to emphasize that the state is provoked by external and situational stimuli, rather than by individual variables (Hidi & Renninger, 2006). The term situational interest has sometimes even been used interchangeably with *interestingness of the situation* to describe the characteristics of tasks and texts that cause the state of interest (e.g., Schraw & Lehman, 2001). Moreover, it has been argued that situational interest occurs primarily in the early phases of interest development, when individual interest is absent or very low (Hidi & Renninger, 2006). Recently, Sansone and Thoman (2005) have therefore suggested that the term interest experience be adopted as a neutral description that does not refer to the genesis of the phenomenon. This term indeed seems better suited to describe the state of interest that occurs while individuals are engaged in learning activities and to convey the idea that the phenomenological state of interest observed may be an outcome of internal and/or external influences.

There is also conceptual overlap between interest experience and the task value component of expectancy-value theory (see Wigfield & Eccles, 2002) in terms of the intrinsic qualities shared by both concepts. Four components of task value can be identified: attainment value, intrinsic value, utility value, and cost. The largest overlaps are with attainment value and intrinsic value (see Eccles, 2005). Our conceptualization thus has much in common with task value as defined in expectancy-value theory, but it is not identical. First, the task value conceptualization is broader than our construct. Second, task values are typically investigated as rather stable beliefs, whereas interest experience is assumed to be a momentary state that may or may not last after an activity has been completed.

### Individual Characteristics That Influence Interest Experience

Interest experience is partly determined by individual characteristics such as gender and prior knowledge that the student brings to the instructional situation. For instance, girls are known to be more interested in certain topics (e.g., living things vs. inanimate objects) than boys (Hidi, 2006). Furthermore, students who lack necessary background knowledge and skills are less likely to experience interest in the classroom. In fact, prior knowledge has been shown to be associated with more interest experience (Alexander, Jetton, & Kulikowich, 1995). Such individual characteristics are fairly stable and unlikely to be altered by situational conditions.

The most discussed of these individual characteristics is the stable personal preference of individual interest (Krapp et al., 1992; Renninger, 2000). *Individual interest* (also referred to as *personal interest*) is defined as a relatively enduring disposition to attend to certain objects, stimuli, or events over time. It relates to specific content and is characterized by positive feelings and

values and accompanied by structured knowledge (Krapp, 2000, 2002b; Krapp et al., 1992; Renninger, 2000; Schiefele, 1991). Content-specific relationships evolve gradually from experience and biological predispositions; accordingly, the configuration of individual interest in different content areas differs across persons (Renninger, 2000). Some researchers regard individual interest as a motivational resource that helps people to cope with unfavorable learning conditions (Katz, Assor, Kanat-Maymon, & Bereby-Meyer, 2006; Silvia, 2006). When working on boring tasks, people with higher individual interest are more likely to engage in interest-enhancing strategies and to transform the activity into something more enjoyable (Sansone, Weir, Harpster, & Morgan, 1992).

Interest theory suggests that the psychological state of interest is automatically triggered when contents are perceived as relevant to one's individual interest. Interest experience is a momentary manifestation of this latent disposition (Krapp, 2002b). Indeed, experimental studies have demonstrated that interest experience can be predicted by individual interest. Particularly in the area of text interest, Ainley, Hillman, and Hidi (2002) found that secondary students with higher individual interest in literature already anticipated higher interest after being presented with the text titles and a few sentences.<sup>1</sup> The authors then used “online recording” methods (Ainley, Hillman, & Hidi, 2002, p. 421) to track participants' interest experiences during the reading process and found that these students continued to experience more interest, showed more persistence, and performed better in a recall test. Similarly, Hidi, Berndorff, and Ainley (2002) showed that general interest in writing is associated with enjoyment of writing activities. In a study involving logic problems, Katz et al. (2006) found that students with higher individual interest in logic questions were more willing to work on such tasks even when given no positive feedback. To date, most studies have been conducted in laboratory conditions, and little is known about the influential effects of prior individual interest in the classroom setting. It seems reasonable to predict that students' individual interest in the academic domain in question is likely to be activated and in turn influences students' interest experience.

At the same time, some findings indicate that individual interest alone is not sufficient to sustain interest experience throughout the learning process. Situational factors may also play a role in influencing students' interest experience over and above individual characteristics. For example, Ainley, Hillman, and Hidi (2002) found that text titles generated different levels of interest experience, regardless of participants' individual interest.

### Situational Factors That Influence Interest Experience

The second source of influence on interest experience is often described as the interestingness of the learning situation in terms of its content, topics, activities, and so forth. These situational factors are naturally assumed to be less stable and more easily manipulated than individual factors. Laboratory research in the area of text

<sup>1</sup> Ainley, Hidi, and Berndorff (2002) labeled this kind of interest that is elicited by a word or paragraph *topic interest*. According to this definition, topic interest is very similar to interest experience. We regard it as an interest experience that occurs in the specific situation of being presented with text and as an outcome of both individual and situational factors.



interest has identified various features of interestingness, such as the coherence, seductiveness, and vividness of the text (Schraw & Lehman, 2001). Mitchell (1993) identified “catch” and “hold” components of instruction that serve to enhance students’ interest state in mathematics classrooms. For instance, using puzzles and computers may initially catch students’ interest, and emphasizing meanings and encouraging students’ involvement can further hold their interest. Bergin (1999) outlined several situational factors that teachers can apply to enhance students’ interest in the classroom, such as hands-on activities, food, games, and puzzles. Consequently, many reform-based programs have endeavored to elicit interest through components such as videos, computer-based lessons, and authentic materials (e.g., Cognition and Technology Group at Vanderbilt, 1991).

The situational factors addressed thus far concern the surface level of learning activities, such as settings and materials, rather than interactional aspects that occur during the process of instruction. In the broader motivation literature, it has been argued that students’ perceptions of classroom instruction, especially of what teachers do and say, are associated with their motivation and behaviors (Ames, 1992; Stefanou, Perencevich, DiCintio, & Turner, 2004; Turner et al., 1998). With its focus on intrinsically motivated and autonomously regulated activities, self-determination theory (SDT) can provide valuable insights into students’ interest experience. Furthermore, SDT provides a rationale for how social and environmental support can promote autonomous behaviors and engagement in learning (Deci, Vallerand, Pelletier, & Ryan, 1991).

According to SDT, intentional behaviors can be motivated by either autonomous or controlled forms of regulation. Autonomous forms of regulation include both intrinsic motivation, or behavior energized by its inherent satisfactions, and identified or integrated forms of extrinsic motivation. Here, the individual identifies with the personal importance of a behavior or assimilates its regulation to the self (Ryan & Deci, 2000a), meaning that regulation is autonomous, volitional, and valued by the self (Deci et al., 1991). Research in SDT has also repeatedly confirmed that autonomously regulated behaviors are characterized by the experience of interest (Deci, 1992; Grolnick & Ryan, 1987). By contrast, behaviors experienced as controlling (e.g., externally regulated or introjected regulations) are typically not associated with either interest or task enjoyment (Ryan & Connell, 1989; Ryan & Deci, 2000a). Accordingly, the level of autonomy support in the classroom is a key factor for understanding students’ interest (Reeve, 2002). Teachers can create an autonomy-supportive climate by attempting to understand students’ feelings and thoughts about learning tasks and by supporting students’ personal growth (Assor, Kaplan, & Roth, 2002). Specifically, autonomy-supportive instructional behaviors include listening, asking questions about students’ wishes, responding to students’ questions and acknowledging the students’ perspective, allowing students to work on their own, using praise as informational feedback, and offering encouragement (Reeve, Bolt, & Cai, 1999; Reeve & Jang, 2006; Williams & Deci, 1996).

Experimental, survey, and longitudinal studies have provided evidence to confirm the positive effects of this type of autonomy support on students’ interest and engagement (Deci et al., 1991; Krapp, 2002a, 2005; Reeve, 2002; Ryan & Deci, 2000b; Trouilloud, Sarrazin, Bressoux, & Bois, 2006; Urdan & Turner, 2005). Findings on the effects of autonomy-supportive instructional be-

haviors in authentic classroom settings are of particular interest for the current investigation (Reeve, 2002; Reeve & Jang, 2006). Students whose teachers are more autonomy oriented in their instructional style (Black & Deci, 2000; Deci, Schwartz, Sheinman, & Ryan, 1981; Ryan & Grolnick, 1986) or are given training in autonomy support (Reeve, Jang, Carrell, Jeon, & Barch, 2004) have been found to show higher intrinsic motivation, more positive emotion, and more active involvement.

In reality, however, the social arrangement of the classroom, with teachers as instructors and students as “receivers,” often leads teachers to neglect students’ need for autonomy and to resort to overly directive or controlling instructional behaviors (Assor, Kaplan, Kanat-Maymon, & Roth, 2005). During instruction, these behaviors include disrupting the students’ natural rhythm of learning (not letting them work at their preferred pace), using directive commands, making “should” statements, and asking controlling questions (see Assor et al., 2002; Reeve & Jang, 2006). These teaching behaviors have been shown to impair students’ sense of autonomy and to hinder intrinsic motivation, engagement, effort, and persistence (Ryan & Grolnick, 1986). From a self-determination perspective, these behaviors are autonomy suppressing in that the teachers fail to support their students’ need for autonomy. Moreover, these behaviors reflect teachers’ attempts to impose a teacher-centered agenda by having an instantaneous impact on students’ behavior and leaving students no room for self-reliant behaviors (Assor et al., 2002; Grolnick & Ryan, 1987). Assor et al. (2005) have demonstrated that controlling instructional behaviors and autonomy support are perceived as distinct aspects of teachers’ behavior. Controlling instructional behaviors, in particular, have a unique effect in inducing the negative emotions of anger and anxiety during the learning process (Assor et al., 2005).

Cognitive autonomy support has recently been proposed as another dimension of autonomy support (Stefanou et al., 2004). Whereas autonomy-supportive climate and controlling instruction focus on social interaction, cognitive autonomy support emphasizes the support provided for students’ engagement in cognitive activities. It has been proposed that students experience a sense of personal control at the cognitive level when teachers explain the purposes of the task at hand and its links to the learning concepts and scaffold students’ understanding by activating prior knowledge or increasing personal relevance (Schraw & Lehman, 2001; Stipek, 1996, 2002; Turner & Meyer, 2004). Despite the claim that cognitive autonomy support “truly leads to psychological investment in learning” (Stefanou et al., 2004, p. 101), there has as yet been little direct empirical investigation of the concept in classrooms. However, instructional research provides indirect support for this claim. For instance, enhancing students’ active cognitive participation has been shown to foster learning and to increase intrinsic interest and enjoyment (Brophy, 1999; Kunter & Baumert, 2006a; Stipek, 1996, 2002; Turner, 2001; Turner et al., 1998; Vermunt & Verloop, 1999). In mathematics classrooms, students show more involvement and positive affect when teachers scaffold learning and transfer responsibilities to students (Turner et al., 1998).

Within the framework of self-determination theory as reviewed above, we can thus identify three features of the instructional process that may affect students’ intrinsic motivation and interest experience: autonomy-supportive climate, controlling behaviors, and cognitive autonomy support. In the present investigation, we



conceptualize all three aspects as situational variables that may differ from lesson to lesson, rather than as features of a stable motivating style (Reeve, 1998). Evidence from experimental designs has repeatedly shown that slight manipulations in autonomy-supportive instruction can significantly affect participants' interest (Deci, Eghrari, Patrick, & Leone, 1994; Reeve, Jang, Hardre, & Omura, 2002; Vansteenkiste, Simons, Lens, Soenens, & Matos, 2005). In the present study, we examine all three instructional aspects simultaneously to determine their distinct influences on students' interest experience in authentic classroom settings.

### The Present Study: Linking Situational and Individual Factors in an Intraindividual Approach

The present study investigates students' interest experience in the real-life classroom environment, examining both the situational and the individual factors that contribute to interest experience in a lesson. An intraindividual approach seemed appropriate to determine the relative effects of situational and individual factors simultaneously. In a repeated-measures design, we followed a group of students over a 3-week period, assessing their interest experience immediately after lessons. Situational factors assumed to vary across lessons were also assessed repeatedly. Multilevel modeling techniques were then applied to analyze intraindividual data and to estimate the effects of situational and individual factors simultaneously. Similar methods have been applied to investigate multilevel data in domains such as individuals' differentiated attachment patterns to different targets (La Guardia, Ryan, Couchman, & Deci, 2000), students' homework effort in different subjects (Trautwein & Lüdtke, 2007), and self-evaluation and mood in diary studies (e.g., Armeli, Carney, Tennen, Affleck, & O'Neil, 2000; Möller & Husemann, 2006).

Specifically, the present study has the following objectives. First, using a repeated-measurement design to assess students' interest experience in real-life classroom settings over time, the study offers unique opportunities to examine intraindividual variation in students' interest experience. Emotional experience has been shown to fluctuate naturally over time (Eid & Diener, 1999); therefore, it seems worthwhile to examine the extent of intraindividual variation in students' interest experience in a structured learning situation such as the classroom.

The second objective was to examine the effects of three aspects of autonomy support as situational factors. On the basis of SDT, we expected variation in students' interest experience across lessons to be predicted by the level of autonomy support experienced in the lesson, not only in terms of autonomy-supportive climate and controlling behaviors but also in terms of cognitive autonomy support. In addition, we explored whether these situational effects applied equally across all students.

Third, beyond situational factors, we examined the individual factors of gender, domain-specific grades, and individual interest. On the basis of interest theory, we expected individual interest as a stable personal preference to be associated with interest experience in situations whose content was perceived to be related. In other words, students' stable subject-specific individual interest was expected to impact their interest experience during lessons in that subject.

To examine the generality of the above hypotheses across different academic domains, we collected student data on three school

subjects. We purposely chose the core school subjects of mathematics, native language instruction (i.e., German), and the second foreign language. At least 4 lesson hours per week were dedicated to each of these compulsory subjects. It is therefore plausible to assume that students' experiences in these lessons are important for their general well-being in school. We chose the second foreign language (rather than English) as a new subject that had only recently been introduced to the students' timetables to contrast with the two old subjects of mathematics and German, in which students had several years' learning experience. This approach allows us to examine whether interest fluctuation is higher in subjects in which students have little prior experience.

## Method

### Participants

Participants were 261 (57% girls) seventh-grade students in Germany. The mean age was 12.3 years ( $SD = 0.5$ ). The vast majority of participants were of European origin ( $> 95\%$ ) and reported speaking German with at least one of their parents (91.2%). Of those who did not speak German at home, 30% reported Turkish and 35% Polish as the primary home language. Students' participation was voluntary and required parental consent; 90% of target students participated. Students were recruited from nine classes in two public *gymnasium* schools in Berlin. *Gymnasium* is the highest track in the three-tier secondary school system in Germany; about one-third of students are enrolled in *gymnasium* schools based on their achievement in elementary school.

### Procedure

The study consisted of a pretest to assess individual characteristics and a lesson-specific repeated measurement phase. In all cases, assessments were administered to classroom units. The pretest was administered in the 4th week of the school year. Lesson-specific repeated measurements began 1 week later and continued for 3 consecutive weeks. The timetables of the nine participating classes were obtained in advance in preparation for the lesson-specific repeated measurements. A maximum of four mathematics lessons, four German (native language instruction) lessons, and four to five second foreign language lessons were scheduled per week. All subjects were taught by different teachers. Data were obtained on all mathematics, German, and second foreign language lessons that took place during the 3-week assessment period. Only lessons that coincided with special events (e.g., class trip) or that did not involve regular instruction (e.g., whole lesson used for an exam) were not assessed.

Lesson-specific measures comprising 33 items were administered at the end of each lesson assessed. The teachers concluded the lesson 3–5 min earlier than scheduled, and research assistants then entered the classroom to administer the lesson-specific questionnaire. Students were instructed to respond on the basis of their experiences during that specific lesson. The first lesson assessed included a 10-min training unit to ensure that all participants understood the questionnaire and procedure. It took students less than 5 min to complete the questionnaires.



## Measures

*Lesson-specific measures (LSM).* Students' interest experience and autonomy-related perceptions in particular lessons were measured using LSM. The same measures were used for mathematics, German, and second foreign language lessons. In designs such as daily diary studies, short instruments are often used to reduce the burden of repeated queries on participants. Some studies have relied on one- or two-item measures (e.g., Birnbaum, Reis, Mikulincer, Gillath, & Orpaz, 2006; Finkel, Burnette, & Scissors, 2007). With a view to the reliability of the measures administered in the present study, we used four to six items to assess each construct measured in LSM (all LSM items are listed in the Appendix).

*Interest experience.* Based on Krapp's (2002b) conceptualization, the interest experience measure comprised an emotion component and a value component. The scale contained a total of five items assessing the emotion component ("The topic was interesting to me," "I liked the topic") and the value component ("The topic was meaningful to me," "I saw what the teacher taught us can be useful in real life," and "It was important to me that I thoroughly understand my class work"). Responses were given on a 6-point scale that ranged from 1 (*disagree strongly*) to 6 (*agree strongly*). We conducted exploratory factor analyses to examine whether the items formed a unidimensional scale for interest experience. The factor analyses<sup>2</sup> showed strong support for a unidimensional factor structure, indicating that different aspects of interest experience form one construct in the LSM. The mean of all five items was therefore calculated separately for each lesson-specific measurement to form the interest experience variable. Cronbach's alpha as an index for internal consistency was calculated separately for each lesson-specific measurement; the mean Cronbach's alpha was .90 (range = .87–.93) for mathematics, .90 (range = .86–.92) for German, and .91 (range = .87–.93) for the second foreign language.<sup>3</sup>

The LSM also assessed the three aspects of autonomy support. We measured these variables from the students' perspective because individual students may experience different amounts of support or perceive the climate and instruction differently, even in the same classroom.

*Perceived autonomy-supportive climate* was measured by a short, six-item version of the Learning Climate Questionnaire (Williams & Deci, 1996). Items were adapted to the lesson situation (e.g., "I felt that my teacher provided me choice and options" and "I felt understood by my teacher"). Responses were given on a 6-point scale that ranged from 1 (*disagree strongly*) to 6 (*agree strongly*).

*Perceived controlling behaviors* were operationalized as overt and inappropriate teaching behaviors that disrupted students' natural rhythm in terms of workload or pacing and that left students little room for self-reliant behaviors. As well as behaviors that have previously been investigated as subscales of autonomy-suppressing behaviors, such as intrusive and overly demanding instructional behaviors, we included items tapping inappropriate instructional behaviors in broader terms (Assor et al., 2002). Four items tapped students' perceptions of these behaviors (e.g., "Our teacher expected split-second answers" and "Our teacher was mean to one of the students").

*Perceived cognitive autonomy support* measured instruction that involves students cognitively and scaffolds their conceptual un-

derstanding with five items (e.g., "More than one student presented their solution to a task" and "Our teacher emphasized the relations between the topics discussed"; Kunter & Baumert, 2006b).

We conducted exploratory factor analyses separately for each assessment of interest experience to examine the factor structure underlying all autonomy-related items. A total of 27 exploratory factor analyses (9 analyses for each of the three subjects) were conducted for the autonomy-related items. As expected, a three-factor solution emerged. Factor analyses with varimax rotation yielded eigenvalues between 4.06 and 7.46 for the first principal component and explained between 36% and 53% of the variance. This component included items pertaining to perceived autonomy-supportive behaviors. The eigenvalues of the second principal components were between 1.80 and 2.39 and explained between 13% and 18% of the variance. This component included items on perceived cognitive autonomy support. With one exception, the eigenvalues of the third principal component were also larger than 1 (range = 0.98–1.7) and explained between 7% and 12% of the variance. This component included items on perceived controlling behavior. The eigenvalues of subsequent principal components were substantially lower, as was the explained variance. The eigenvalues of the fourth principal component were between .82 and 1.14, and the explained variance was between 6% and 9%. This component included one or two items measuring controlling behaviors. Overall, 22 of the 27 (81%) exploratory factor analyses yielded three principal components larger than 1. The other five factor solutions yielded four principal components larger than 1. Based on the factor solutions, three variables of lesson-specific autonomy-related perceptions were thus calculated for each lesson-specific repeated measurement: perceived autonomy-supportive climate, perceived controlling behaviors, and perceived cognitive autonomy support. These three scales were moderately correlated with each other: Larger correlations were found between perceived autonomy-supportive climate and perceived cognitive autonomy support ( $r \leq .35$ ); the correlations with perceived controlling behaviors were below .11 (see the correlations above the diagonals in Table 1). Cronbach's alpha was calculated for

<sup>2</sup> Exploratory factor analyses were conducted separately for each measurement at the lesson level. Nine exploratory factor analyses were conducted for each school subject (only measurement points containing data from at least 100 participants were analyzed). The same pattern emerged for all three subjects. In mathematics lessons, only the first principal component had eigenvalues larger than 1 (range = 3.7–4.4), with explained variance ranging between 62% and 73%. The eigenvalues (range = 0.55–0.93) and explained variance (below 15%) for the second principal component were much lower. Likewise, in German lessons, only the first principal component had eigenvalues larger than 1 (range = 3.5–4.4), with explained variance ranging between 58% and 73%. The eigenvalues (range = 0.56–0.95) and explained variance (below 16%) for the second principal component were much lower. Finally, in the second foreign language, only the first principal component had eigenvalues larger than 1 (range = 3.7–4.7), with explained variance ranging between 61% and 78%. The eigenvalues (range = 0.56–0.77) and explained variance (below 13%) for the second principal component were much lower.

<sup>3</sup> Applying the same rule as in the exploratory factor analyses, Cronbach's alphas were calculated separately for measurement points on the lesson level containing data from at least 100 participants. In total, nine Cronbach's alphas were calculated for each school subject.

Table 1  
Correlations Among Lesson-Specific Measures and Individual Characteristics

Variable	Mathematics				German				2-language			
	1	2	3	4	1	2	3	4	1	2	3	4
Lesson-specific measures												
1. Interest experience	—	.35**	-.12**	.34**	—	.36**	-.09**	.32**	—	.39**	-.19**	.33**
2. Autonomy-supportive climate	.57**	—	-.05*	.34**	.63**	—	-.04	.35**	.58**	—	-.10**	.34**
3. Controlling behaviors	-.19*	.18	—	.07**	.04	.20**	—	.11**	-.12*	.18**	—	.08**
4. Cognitive autonomy support	.66**	.68**	.01	—	.54**	.71**	.26**	—	.55**	.70**	.16**	—
Individual characteristics												
Individual interest	.51**	.26**	-.11	.26**	.42**	.27**	.01	.16**	.52**	.27**	-.10	.21**
Subject-specific school grade	.13*	.03	-.04	.07	-.02	.03	-.11	-.02	—	—	—	—

Note. For the lesson-specific measures, correlations above the diagonals represent within-person correlations and correlations below the diagonals represent between-person correlations. Dashes indicate that no data were available. 2-Language = second foreign language.

\*  $p < .05$ . \*\*  $p < .01$ .

each measurement at the lesson level as an index of internal consistency. For mathematics, German, and the second foreign language, respectively, the mean  $\alpha = .92$  (range = .86–.96), .92 (range = .86–.94), and .92 (range = .84–.94) for perceived autonomy-supportive climate; .66 (range = .59–.73), .64 (range = .59–.70), and .69 (range = .60–.75) for perceived controlling behaviors; and .76 (range = .62–.85), .76 (range = .58–.83), and .76 (range = .65–.84) for perceived cognitive autonomy support.

### Measures of Individual Characteristics

Individual characteristics such as individual interest, school grades, and gender were assessed at the pretest. A 7-item scale was used to assess individual interest in mathematics, German, and the second foreign language (Ramm et al., 2006). These items (e.g., “I do [subject matter], because it is fun for me” and “For me, [subject matter] is personally important”) were based on earlier work (e.g., Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005). Students responded to each question in separate columns for mathematics, German, and the second foreign language. Responses were given on a 6-point scale anchored by the end points 1 (*disagree strongly*) and 6 (*agree strongly*). Cronbach’s alphas for mathematics, German, and the second foreign language were .92, .87, and .91, respectively. Students were also asked to report the mathematics and German grades they had been awarded on their sixth-grade report cards (i.e., at the end of elementary school). Elementary school grade for second foreign language was not available because the students had not been introduced to this subject until the seventh grade. The German grading system ranges from 1–6, with smaller numbers indicating better performance. To maintain consistency with the other scales, we reverse coded the grades such that a higher value indicates better performance. Gender was coded 0 for female and 1 for male.

### Statistical Analysis

**Analyzing hierarchically structured data.** The repeated measures design results in a two-level hierarchical structure with LSMs nested within each individual student. At the within-student level (Level 1), we tested a number of repeated LSMs collected for each student over the 3-week period. At the between-student level

(Level 2), we assessed variations in student characteristics including gender, individual interests, and school grades.

Hierarchical linear modeling (HLM; Raudenbush & Bryk, 2002) is capable of handling hierarchically structured data that violate the assumption of independence. Furthermore, HLM identifies variance in multiple levels of the data—in our case, at the within- and between-students levels. Moreover, HLM allows simultaneous estimation of effects of predictors from different levels. Detailed descriptions of the models are provided with the results. All analyses were conducted using the program HLM (version 6; Raudenbush, Bryk, Cheong, & Congdon, 2004).

**Treatment of missing values.** Our data set contained three types of missing data: first, missing responses to some of the items in the LSMs; second, missing responses to measures of individual characteristics in the pretest; and third, missing data caused by students’ absence in some of the lesson-specific assessments. Because HLM allows unequal numbers of observations for each individual, missing data of the third type were not a problem. The first two types of missing data were not considered problematic because the rate for each was less than 3%. Different approaches have been shown to produce satisfactory results when the missing rate is below 5% (Graham, Cumsille, & Elek-Fisk, 2003). In the present study, a multiple imputation procedure was chosen (Peugh & Enders, 2004). All available data from the lesson-specific and pretest measures were used to estimate the missing values, including demographic and motivational variables that were not used in the present investigation. The NORM software (Version 2.03; see Schafer & Graham, 2002) was used to generate five data sets in which missing data were replaced by estimated values. All HLM analyses were conducted with the five complete data sets, and combined results are reported (Schafer & Graham, 2002). The descriptive statistics reported below are based on the raw data set.

## Results

### Descriptive Analyses

A total of 6,468 completed lesson-specific questionnaires were obtained; students participated in an average of 8.4 measurements



(range = 5–11) in mathematics lessons, 8.0 measurements (range = 2–10) in German lessons, and 8.4 measurements (range = 6–11) in second foreign language lessons.

For exploratory purposes, mean values of lesson-specific measurements were calculated for each individual student for mathematics, German, and the second foreign language separately. Descriptive statistics for these person-mean LSMs are presented in Table 2. In general, students' person-mean scores on interest experience tended toward the positive end of the 6-point scale. In terms of autonomy support, within-person means for autonomy-supportive climate and cognitive autonomy support were around the midpoint of the scale, whereas person-mean scores on perceived controlling behaviors were low.

Intercorrelations among LSMs and individual characteristics are presented in Table 1. Two types of correlations were computed for the LSMs: between-person correlations and within-person correlations. Between-person correlations were calculated using the mean scores for each student. As shown under the diagonals in the top panel of Table 1, person-mean interest experience was moderately associated with person-mean autonomy-supportive climate ( $r \geq .57$ ) and person-mean cognitive autonomy support ( $r \geq .54$ ), indicating that students who reported higher average interest experience also reported higher average autonomy-supportive climate and cognitive autonomy support. Findings on the association between person-mean interest experience and person-mean controlling behaviors were mixed across three subjects.

Regarding the relations to individual characteristics, students' person-mean interest experience correlated significantly with individual interest in the corresponding academic domain as measured in the pretest ( $.42 \leq r \leq .52$ ). Smaller but significant correlations were found between autonomy-supportive climate and perceived cognitive autonomy support and individual interest ( $.16 \leq r \leq .27$ ). No significant associations were found between perceived controlling behaviors and individual interest.

Second, within-person correlations were calculated using all lesson-specific data. These within-person correlations indicate the correlation structure across lessons for individual students. In order to partial out the between-person variance, each student's person-

mean value was subtracted from his/her raw scores, such that every student had the new mean value of zero. Next, all lesson-specific scores from students were used to calculate the correlations. As shown above the diagonals in the top panel of Table 1, interest experience correlated significantly with all three autonomy-related perceptions. It has been suggested that the within-person correlation structure may reveal different results from the between-person correlation structure (Michela, 1990). In our case, the within-person correlations showed that students reported less interest in lessons where they perceived more teacher control, but this association was not always significant across school subjects in the between-person correlations. In the next section, we examine the within-student and between-student variance simultaneously with HLM models.

Examining Intraindividual Variation in LSMs

We used an unconditional model to examine variance in interest experience at the within-student and between-student level. If a student tends to produce similar responses across lessons, the proportion of variance at the within-student level will be low. Separate models were calculated for mathematics, German, and the second foreign language. The results show that the proportion of variance at the within-student level was substantial: 36% in mathematics, 45% in German, and 36% in the second foreign language. Therefore, our findings supported the hypothesis of lesson-to-lesson variation in students' interest experience.

In addition, unconditional HLM models were applied to analyze the variance components of the three autonomy-related LSMs. Proportions of within-student variance were 36%–38% for perceived autonomy-supportive climate, 52%–58% for controlling behaviors, and 44%–50% for cognitive autonomy support. The sizeable variances in both interest experience and the three predictors justified our further attempts to predict interest experience by reference to three autonomy-related perceptions at the within-student level.

We begin with a detailed description of the HLM analyses incorporating predictors from both levels. The analyses served two main objectives. First, we examined whether students' interest experience during lessons could be predicted by situational factors of the lesson (i.e., within-student level predictors). Second, we investigated whether individual differences in interest experience could be predicted by student characteristics (i.e., between-student level predictors). Models were fitted using restricted maximum likelihood estimation. The results for within-student and between-student level predictors are described in the following sections.

First, at the within-student level, we investigated how interest experience is predicted by the three autonomy-related perceptions using the regression equation below.

Interest<sub>ij</sub> =  $\pi_{0i}$  +  $\pi_{1i}$ Autonomy +  $\pi_{2i}$ Control +  $\pi_{3i}$ Cognitive +  $\epsilon_{ij}$ ,

where Interest<sub>ij</sub> is the interest experience of the *i*th student in the *j*th lesson;  $\pi_{0i}$  represents the intercept of the *i*th student;  $\pi_{1i}$ ,  $\pi_{2i}$ ,  $\pi_{3i}$  represent the regression coefficients for the three autonomy-related perceptions; and  $\epsilon_{ij}$  is random within-person error. All three predictors were group-mean centered (here, person centered). This procedure tests specifically whether interest experience is ex-

Table 2  
Person-Mean Lesson-Specific Measures and Measures of Individual Characteristics

Variable	Mathematics		German		2-language	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Lesson-specific measures						
PM Interest experience	3.94	1.18	3.95	1.09	4.11	1.15
PM Autonomy-supportive climate	3.05	1.13	3.16	1.15	3.15	1.12
PM Controlling behavior	1.95	0.73	1.71	0.62	1.92	0.70
PM Cognitive autonomy support	3.46	1.04	3.11	1.01	3.14	1.03
Individual characteristics						
Individual interest	3.90	1.34	3.82	1.06	4.57	1.11
Subject-specific school grade	4.84	0.77	4.99	0.55	—	—

Note. All items were rated on a 6-point scale with 6 as the highest value. Dashes indicate that no data were available. PM = person-mean; 2-language = second foreign language.

plained by different autonomy-related perceptions relative to an individual student's own baseline. Alternative procedures (e.g., no centering or grand-mean centering) are less suitable for the present investigation, because the results might be caused partly by students' different baselines. The present approach allows effects purely at the within-student level to be disentangled.

Turning to the between-student level, the parameters are further modeled by individual characteristics as follows:

$$\pi_{0i} = \gamma_{00} + \gamma_{01}\text{Male} + \gamma_{02}\text{IndInt} + \gamma_{03}\text{Grade} + \zeta_{0i},$$

$$\pi_{1i} = \gamma_{10},$$

$$\pi_{2i} = \gamma_{20},$$

$$\pi_{3i} = \gamma_{30}.$$

The first equation predicts the intercept  $\pi_{0i}$ —the average interest experience of the  $i$ th student—derived from a Level 1 equation. Here, it is predicted by the individual characteristics of gender, individual interest, and school grade. The  $\gamma_{00}$  represents the grand mean of the sample accounting for the effects of gender on the students' mean interest experience ( $\gamma_{01}$ ), the effects of individual interest and prior school grades ( $\gamma_{02}$  and  $\gamma_{03}$ ). The  $\zeta_{0i}$  reflects random error at the between-student level. The next three equations indicate that the within-student level regression parameters  $\pi_{1i}$ ,  $\pi_{2i}$ , and  $\pi_{3i}$  are treated as fixed effects and thus only predicted by the intercepts  $\gamma_{10}$ ,  $\gamma_{20}$ , and  $\gamma_{30}$ . All continuous variables were z-standardized prior to multilevel modeling.

### Predicting Interest Experience With Situational Factors

Model 1 tested the effects of the three within-student level predictors that are presented in Table 3. The three school subjects are modeled separately. The hypothesized associations between the three autonomy-related perceptions and the change in students'

interest from lesson to lesson were all supported. As shown, students experienced more interest in lessons where they perceived (relative to their own baseline) a more autonomy-supportive climate, more cognitive autonomy support, and less controlling behavior. In mathematics lessons, for example, autonomy-supportive climate and cognitive autonomy support had moderate effects, with  $B$  values of .25 and .23, respectively. In other words, one standard deviation increase in autonomy-supportive climate and cognitive autonomy support was associated with one quarter of a standard deviation increase in interest experience. The effect of controlling behavior was statistically significant but smaller, at  $B = -.10$ . A consistent pattern of results was found across all three subjects. Overall, over 19% of the variance at the within-student level was explained by the three autonomy-related perceptions.

### Predicting Interest Experience With Between-Student-Level Factors

Turning to the between-student level, we examined whether the differences in students' average interest experience could be predicted by the individual characteristics of gender, individual interest, and school performance. As shown in Table 3 (Model 1), we found a unique effect of individual interest among the three individual characteristics, with coefficients between .35 and .46 across the three subjects. Interest experience was thus significantly predicted by individual interest in the subject. There were no significant effects of gender or school grades on interest in the subject. Overall, the individual predictors explained 27% of the variance at the between-student level in mathematics, 19% in German, and 27% in the second foreign language.

One open question remains regarding the effects of individual characteristics. Is it possible that some students reported higher average interest experience because lessons were constantly more

Table 3  
Using Hierarchical Linear Modeling to Predict Interest Experience

Predictor	Mathematics				German				2-Language			
	Model 1		Model 2		Model 1		Model 2		Model 1		Model 2	
	<i>B</i>	<i>SE</i>	<i>B</i>	<i>SE</i>	<i>B</i>	<i>SE</i>	<i>B</i>	<i>SE</i>	<i>B</i>	<i>SE</i>	<i>B</i>	<i>SE</i>
Within-student level												
Autonomy-supportive climate	.25***	.03	.25***	.03	.31***	.03	.31***	.03	.28***	.03	.28***	.03
Controlling behaviors	-.10***	.02	-.10***	.02	-.10***	.02	-.10***	.02	-.14***	.02	-.14***	.02
Cognitive autonomy support	.23***	.03	.23***	.03	.23***	.03	.23***	.03	.22***	.03	.22***	.03
Between-student level												
Gender <sup>a</sup>	-.13	.09	.02	.07	.01	.09	.07	.07	.04	.09	.03	.07
Individual interest	.46***	.05	.28***	.04	.35***	.04	.24***	.04	.44***	.04	.29***	.04
Subject-specific school grade	-.06	.04	-.04	.03	-.08	.04	-.06	.04	—	—	—	—
PM autonomy-supportive climate			.19***	.06			.34***	.06			.26***	.05
PM controlling behaviors			-.15***	.03			-.09**	.03			-.15***	.04
PM cognitive autonomy support			.35***	.06			.16**	.06			.23***	.05
Explained variance												
Lesson level	.19		.19		.19		.17		.23		.23	
Individual level	.27		.63		.19		.52		.27		.57	

Note. Lesson-level predictors were group-mean centered. Dashes indicate that no data were available. PM = person-mean; 2-Language = second foreign language;  $B$  = unstandardized regression coefficient resulting from hierarchical linear modeling analyses;  $SE$  = standard error.

<sup>a</sup> Gender coded: girl = 0 and boy = 1.

\*\*  $p < .01$ . \*\*\*  $p < .001$ .



autonomy supportive for them? Group-mean centering had removed the information on students' baseline lesson perceptions. Therefore, in order to examine the effect of individual characteristics after controlling for the general level of the three autonomy-related lesson perceptions, we followed the recommendation by Kreft and de Leeuw (1998, p. 108) and added the subtracted mean structure of the three group-mean centered predictors back into the models as between-student predictors. Another set of HLM models (Model 2) was thus specified that included averaged within-student level predictors at the between-student level.

In Model 2, the significance pattern of individual characteristics effects remained the same for all school subjects. The effects of individual interest decreased but remained statistically significant; the effects of gender and school grade remained nonsignificant. The regression coefficients for individual interest ranged between .24 and .29 in three school subjects. The coefficients indicated that, even when the overall level of the three autonomy-related lesson perceptions was controlled, students' individual interest was still related to interest experience. The overall variance explained by between-student level predictors increased by at least 30% when the person means of three autonomy-related lesson perceptions were included.

#### *Examining Variation in the Effects of Situational Factors Across Students*

Thus far, the HLM models have assumed that the effects of within-student level predictors are the same for all students. To explore whether the effects of autonomy-supportive climate, controlling behavior, and cognitive autonomy support apply equally to all students, we conducted an additional set of so-called random effect models. In between-student-level HLM equations, random components were specified for  $\pi_{1i}$ ,  $\pi_{2i}$ , and  $\pi_{3i}$ . In other words, a total of nine random components were examined for all school subjects. The results of the random effect model showed that all nine variance components for within-student-level regression coefficients were significant, indicating that the effects of autonomy-supportive climate, cognitive autonomy support, and perceived controlling behaviors on interest experience vary substantially between students. To gain insight into the magnitude of variation, we followed the recommendation of Raudenbush and Bryk (2002, p. 78), and calculated the 95% plausible value range of the slopes using information on the average slope and the slope standard deviation. The largest range of effect was found for autonomy-supportive climate in mathematics lessons: The average slope was 0.27 and the standard deviation was 0.26. It follows that the effect ranged between  $0.27 - 2 \times 0.26 = -.25$  and  $0.27 + 2 \times 0.26 = 0.79$  for approximately 95% of the students in the present study. The smallest range of effect was found for perceived controlling behaviors in mathematics lessons, at between  $-0.36$  and  $0.16$ .

These results revealed meaningful difference across students that warrants further examination. We ran exploratory models in an attempt to explain this variation in terms of students' individual characteristics. Specifically, we examined whether these differences in the slopes across students were explained by gender, individual interest, or school grades. Two statistically significant results emerged. In mathematics lessons, the effect of perceived controlling behaviors was moderated by individual interest ( $B = -.05$ ,  $p = .029$ ), and in second foreign language lessons, the effect

of autonomy-supportive climate was moderated by individual interest ( $B = -.10$ ,  $p = .002$ ). These findings indicated that students with higher individual interest were less affected by the controlling behaviors of their mathematics teachers or by an autonomy-supportive climate in second foreign language lessons. Overall, rather little of the difference across students was explained by the three individual characteristics.

## Discussion

The multilevel analyses conducted to investigate students' interest experience in authentic learning situations revealed three main findings. First, there was substantial intraindividual variation in students' interest experience in day-to-day classroom learning. Second, the pattern of variation was predicted by the situational factors of autonomy supportive climate, controlling behaviors, and cognitive autonomy support in lessons. The effects of situational factors were consistent across the three subjects investigated. Third, students who started with higher domain-specific individual interest had higher interest experience in average lessons.

#### *Interest Experience in Lessons: Intraindividual Variability*

The present study is among the first to investigate how students' interest experience emerges and endures in the classroom over a relatively short period of time. Over three weeks, in an average of eight lessons, variance in interest experience at the within-student level accounted for between 36% and 44% of the overall variance. Whereas between-student variance indicates that interest experience differs from one student to another, within-student variance indicates that interest experience also differs within students from one lesson to another. Moreover, the amount of intraindividual variation observed was very similar across the three subjects examined, although it was slightly higher in German lessons (44%). The extent of variation in the structured learning condition of the classroom thus seems to be similar across domains. To counter the possibility that such within-student variation is mainly due to measurement error, we used a 5-item scale with good internal reliability in the present study. Furthermore, the finding that intraindividual variability covaries with lesson perceptions also indicates that the pattern of results is more meaningful than random fluctuation.

The psychological state of interest is not a fixed entity. Even individuals who generally enjoy reading are more interested in some texts or topics than others (Ainley, Hidi, & Berndorff, 2002; Ainley, Hillman, & Hidi, 2002). To date, most research on interest variation has been carried out in laboratory settings and has focused on text-based processing. The present findings confirm that variation in interest state is observable not only in laboratory settings but also in authentic classrooms across a wide range of topics and activities. The students in our sample, who had six years of classroom learning experience and had already attended numerous lessons in certain subjects, nevertheless experienced some lessons as more interesting and engaging than others. These results challenge the beliefs of some teachers that many students are just not interested and cannot be motivated. The finding that students are sensitive to the learning conditions afforded by the teacher is encouraging for teachers. At the same time, the challenge remains

of how to create an interesting and motivating learning environment.

### *Instructional Features and Interest Experience*

Which instructional features make a lesson substantively different from others? SDT proposes that individuals have a basic need to feel self-determined or autonomous, and that teachers' autonomy support during instruction can facilitate satisfaction of this psychological need (Reeve, 2002). Teachers' autonomy-supportive behaviors have previously been shown to generate more engagement and effort among students (Reeve et al., 2002, 2004). Our study extends these beneficial effects to include students' interest experience. Lessons in which students perceive the teacher taking their perspective and understanding what they want (i.e., an autonomy-supportive climate) are associated with higher interest experience; in contrast, lessons in which teachers disrupt students' natural learning rhythms and do not allow time for reflection (i.e., controlling behaviors) were associated with lower interest experience. The factor solution and small within-person correlations showed that teachers' controlling behavior is a distinct construct from autonomy-supportive teacher behaviors. Moreover, teachers' controlling instructional behaviors seem to be associated with negative student emotions, such as anxiety in the classroom (Assor et al., 2005; Tsai, Kunter, Lüdtke, & Trautwein, 2007). Following the conceptualization of Assor et al. (2002), we measured the construct of controlling behavior in broader terms than has been done in many previous studies. Our items also covered inappropriate instructional behaviors that have been proposed to undermine students' sense of autonomy. Whether the level of autonomous behavior mediates the relation between controlling behavior and interest experience will have to be investigated directly in future research.

Moreover, beyond the social interaction aspect of autonomy-supportive climate, we found a distinctive effect of *cognitive autonomy support*. Stefanou et al. (2004) hypothesized that cognitive autonomy support gives students an enhanced sense of control, particularly during engagement with cognitive activities (e.g., problem solving). Cognitive autonomy is supported when the teacher explicitly explains the aim of tasks and activates students' prior knowledge during their implementation. Within SDT, this kind of practice is discussed in terms of the rationale it provides students for grasping and assimilating the value and meaning of an activity, which has been experimentally shown to facilitate more interested engagement with a task (e.g., Deci et al., 1994). That being said, cognitive autonomy differs from a perceived expectation or standard in the classroom to perform well or to learn harder (cf. academic press for understanding, Middleton & Midgley, 2002). Our study provides empirical evidence within actual classroom settings of the engaging effect of cognitive autonomy support as proposed by Stefanou et al. (2004) by demonstrating the effects of cognitive autonomy support over and above those of autonomy-supportive climate and controlling behaviors. Our findings indicate that lessons in which students' prior knowledge and conceptual understanding are activated and the aims of tasks are transparent to students are associated with enjoyment.

Furthermore, the present findings indicate that perceived autonomy support may depend on what teachers say and do in the classroom and that it is less stable than teachers' individual char-

acteristics. Relative to each student's own baseline, some lessons were perceived as more autonomy supportive than others. Therefore, it might be possible to enhance interest in lessons by providing teachers with training in autonomy-supportive teaching (Reeve, 1998).

The consistent findings across the three subject domains indicate that the beneficial effects of autonomy support are quite general and probably apply to other subjects as well. Nevertheless, the significant variation in the effects found among students indicates that some students seem to react more to teachers' autonomy support than others. To understand how autonomy support can benefit different types of students, the underlying processes moderating these effects need to be further investigated.

### *Individual Interest and Interest Experience*

Interest theory predicts that people who have a stable preference for a certain subject domain or content will seek out related activities and that they will enjoy and value opportunities to reengage with relevant contents. As Krapp (2002b; Krapp et al., 1992) argued, individual interest can be conceptualized as a stable person-object relation that, once developed, will influence the quality of further interactions. The present findings clearly show that students' individual interest in a school subject, measured at the beginning of secondary school, significantly predicts their interest experience in the respective lessons over a 3-week period. It is surprising that effects of a similar magnitude were also found for a newly introduced subject (i.e., second foreign language). We had not expected the person-object relation to be as well developed in this context. Whether individual interest in this subject is a function of prior experience outside school or inferred from other related domains (e.g., general interest in foreign language) remains to be investigated.

Our findings indicate that individual interest can be regarded as a motivational resource for students in everyday classroom learning situations (Katz et al., 2006; Sansone et al., 1992). Although lessons may not always coincide with their preferences, and external support may differ from one day to the next, students with higher individual interest in a subject are more likely to have positive learning experiences in the respective lessons. Elementary school grades did not prove to be associated with interest in secondary classrooms. Nevertheless, the association found between individual interest in the subject and interest experience was not perfect. An interest state is rarely triggered by individual interest alone, even when individual interest is very strong (Hidi & Renninger, 2006). A comprehensive account of interest state needs to take other motivational resources and situational factors into account.

### *Toward the Integration of Situational and Individual Factors in Interest Theory*

The present study showed that interest experience as a momentary psychological state is influenced by both situational factors and individual characteristics. Although similar approaches have already been proposed in the interest literature (e.g., Hoffmann, Krapp, & Renninger, 1998), most researchers have investigated either the situational aspect (i.e., situational interest) or the dispositional aspect (i.e., individual interest), meaning that one source of



influence has generally been neglected. For instance, most studies on situational interest have overlooked the influence of individual characteristics (see Schraw & Lehman, 2001, for a review). Working within this framework, it would also be possible to examine the relative importance of other situational factors proposed in the literature with respect to further individual characteristics or motivational resources. Moreover, the interaction between the two sources of influence (e.g., for which persons certain situational factors are effective) could be further investigated.

Broadly speaking, the intraindividual variability observed in interest experience across situations is in line with state-trait theories from personality and social psychology (Eid & Diener, 1999; Steyer, Schmitt, & Eid, 1999), which predict emotion and behavior to have both state and trait aspects. Students' interest and other emotional experiences should be no exception. In fact, much of literature has referred implicitly to the state-trait distinction. It remains for empirical research to examine the magnitude of intraindividual variability in interest across time and situations and to determine whether this intraindividual variability differs across educational environments, phases of interest development, or different trait levels of individual interest.

### Limitations and Future Research

The results of the present study are in line with hypotheses derived from SDT and interest theory, but several empirical limitations warrant discussion. First, the effects of situational factors (e.g., autonomy-supportive climate) established with multilevel analysis are correlational in nature; therefore, inferences cannot be drawn on the causal direction of the effects of the three situational factors on interest experience. For instance, it is possible that teachers pay more attention to students who show interest during the lesson and give them more positive feedback, and thus they are perceived by these students to be autonomy-supportive. Additionally, there is a possibility of third-variable explanations. For instance, external pressures (e.g., impending examinations) may prime negative mood in students (and perhaps also the teacher) and bias both interest experience and the learning situation in a negative direction. To address this limitation, future studies might investigate the effects in a sequence of experimental learning conditions that allow situational factors to be controlled or manipulated.

Second, data on both interest experience and situational factors of the lesson were obtained from students' self-reports. Some relations may therefore be overestimated due to shared method variance. Further research might address this issue by using multiple sources of information (e.g., teacher reports, third-person observations, analysis of instructional tasks) to provide more objective perceptions of instruction (Kunter & Baumert, 2006b).

A third limitation relates to the generalization of the results. Participants in the present study were sampled from the same grade and the same academic track of the three-tier German secondary educational system. To establish their generalizability, the effects observed in the present study need to be replicated in more heterogeneous samples, and at different points in student development. Replication of the present results in samples of different ages and cultural backgrounds would support the claim of self-determination theory that autonomy support is universally beneficial for all individuals.

### References

- Ainley, M., Hidi, S., & Berndorff, D. (2002). Interest, learning, and the psychological processes that mediate their relationship. *Journal of Educational Psychology, 94*, 545–561.
- Ainley, M., Hillman, K., & Hidi, S. (2002). Gender and interest processes in response to literary texts: Situational and individual interest. *Learning & Instruction, 12*, 411–428.
- Alexander, P. A., Jetton, T. L., & Kulikowich, J. M. (1995). Interrelationship of knowledge, interest, and recall: Assessing a model of domain learning. *Journal of Educational Psychology, 87*, 559–575.
- Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*, 261–271.
- Armeli, S., Carney, M. A., Tennen, H., Affleck, G., & O'Neil, T. (2000). Stress and alcohol use: A daily process examination of the stressor-vulnerability model. *Journal of Personality and Social Psychology, 78*, 979–994.
- Assor, A., Kaplan, H., Kanat-Maymon, Y., & Roth, G. (2005). Directly controlling teacher behaviors as predictors of poor motivation and engagement in girls and boys: The role of anger and anxiety. *Learning and Instruction, 15*, 397–413.
- Assor, A., Kaplan, H., & Roth, G. (2002). Choice is good, but relevance is excellent: Autonomy-enhancing and suppressing teacher behaviours predicting students' engagement in schoolwork. *British Journal of Educational Psychology, 72*, 261–278.
- Bergin, D. A. (1999). Influences on classroom interest. *Educational Psychologist, 34*, 87–98.
- Birnbaum, G. E., Reis, H. T., Mikulincer, M., Gillath, O., & Orpaz, A. (2006). When sex is more than just sex: Attachment orientations, sexual experience, and relationship quality. *Journal of Personality and Social Psychology, 91*, 929–943.
- Black, A. E., & Deci, E. L. (2000). The effects of instructors' autonomy support and students' autonomous motivation on learning organic chemistry: A self-determination theory perspective. *Science Education, 84*, 740–756.
- Brophy, J. (1999). Toward a model of the value aspects of motivation in education: Developing appreciation for particular learning domains and activities. *Educational Psychologist, 34*, 75–85.
- Cognition and Technology Group at Vanderbilt. (1991). The Jasper Series as an example of anchored instruction: Theory, program description, and assessment data. *Educational Psychologist, 27*, 291–315.
- Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety: Experiencing flow in work and play*. San Francisco: Jossey-Bass.
- Deci, E. L. (1992). The relation of interest to the motivation of behavior: A self-determination theory perspective. In A. Renninger, S. Hidi & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 43–70). Hillsdale, NJ: Erlbaum.
- Deci, E. L., Eghrari, H., Patrick, B. C., & Leone, D. R. (1994). Facilitating internalization: The self-determination theory perspective. *Journal of Personality, 62*, 119–142.
- Deci, E. L., Schwartz, A. J., Sheinman, L., & Ryan, R. M. (1981). An instrument to assess adults' orientations toward control versus autonomy with children: Reflections on intrinsic motivation and perceived competence. *Journal of Educational Psychology, 73*, 642–650.
- Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation and education: The self-determination perspective. *Educational Psychologist, 26*, 325–346.
- Eccles, J. S. (2005). Subjective task value and the Eccles et al. model of achievement-related choices. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 105–121). New York: Guilford Press.
- Eid, M., & Diener, E. (1999). Intraindividual variability in affect: Reliability, validity, and personality correlates. *Journal of Personality and Social Psychology, 76*, 662–676.
- Finkel, E. J., Burnette, J. L., & Scissors, L. E. (2007). Vengefully ever

- after: Destiny beliefs, state attachment anxiety, and forgiveness. *Journal of Personality and Social Psychology*, 92, 871–886.
- Graham, J. W., Cumsille, P. E., & Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (pp. 87–114). New York: Wiley.
- Grolnick, W. S., & Ryan, R. M. (1987). Autonomy in children's learning: An experimental and individual difference investigation. *Journal of Personality and Social Psychology*, 52, 890–898.
- Hidi, S. (2006). Interest: A unique motivational variable. *Educational Research Review*, 1, 69–82.
- Hidi, S., & Anderson, V. (1992). Situational interest and its impact on reading and expository writing. In A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 3–25). Hillsdale, NJ: Erlbaum.
- Hidi, S., Berndorff, D., & Ainley, M. (2002). Children's argument writing, interest and self-efficacy: An intervention study. *Learning and Instruction*, 12, 429–446.
- Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*, 41, 111–127.
- Hoffmann, L., Krapp, A., & Renninger, K. A. (Eds.). (1998). *Interest and learning*. Kiel, Germany: IPN.
- Katz, I., Assor, A., Kanat-Maymon, Y., & Bereby-Meyer, Y. (2006). Interest as a motivational resource: Feedback and gender matter, but interest makes the difference. *Social Psychology of Education*, 9, 27–42.
- Köller, O., Baumert, J., & Schnabel, K. U. (2003). Secondary school as a constraint for adolescent development. In U. M. Staudinger & U. Lindenberger (Eds.), *Understanding human development: Dialogues with lifespan psychology* (pp. 449–461). Dordrecht, Netherlands: Kluwer Academic.
- Krapp, A. (2000). Interest and human development during adolescence: An educational-psychological approach. In J. Heckhausen (Ed.), *Motivational psychology of human development: Developing motivation and motivating development*. (pp. 109–129). New York: Elsevier Science.
- Krapp, A. (2002a). An educational-psychological theory of interest and its relation to SDT. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 405–427). New York: University of Rochester Press.
- Krapp, A. (2002b). Structural and dynamic aspects of interest development: Theoretical considerations from an ontogenetic perspective. *Learning and Instruction*, 12, 383–409.
- Krapp, A. (2005). Basic needs and the development of interest and intrinsic motivational orientations. *Learning and Instruction*, 15, 381–395.
- Krapp, A., Hidi, S., & Renninger, K. A. (1992). Interest, learning, and development. In A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 3–25). Hillsdale, NJ.
- Kreft, I. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage.
- Kunter, M., & Baumert, J. (2006a). Linking TIMSS to research on learning and instruction: A re-analysis of the German TIMSS and TIMSS video data. In S. J. Howie & T. Plomp (Eds.), *Contexts of learning mathematics and science: Lessons learned from TIMSS* (pp. 335–351). London and New York: Routledge.
- Kunter, M., & Baumert, J. (2006b). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, 9, 231–251.
- La Guardia, J. G., Ryan, R. M., Couchman, C. E., & Deci, E. L. (2000). Within-person variation in security of attachment: A self-determination theory perspective on attachment, need fulfillment, and well-being. *Journal of Personality & Social Psychology*, 79, 367–384.
- Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, 76, 397–416.
- Michela, J. L. (1990). Within-person correlational design and analysis. In C. Hendrick & M. S. Clark (Eds.), *Research methods in personality and social psychology* (pp. 279–311). Thousand Oaks, CA: Sage.
- Middleton, M. J., & Midgley, C. (2002). Beyond motivation: Middle school students' perceptions of press for understanding in math. *Contemporary Educational Psychology*, 27, 373–391.
- Mitchell, M. (1993). Situational interest: Its multifaceted structure in the secondary school mathematics classroom. *Journal of Educational Psychology*, 85, 424–436.
- Möller, J., & Husemann, N. (2006). Internal comparisons in everyday life. *Journal of Educational Psychology*, 98, 342–353.
- Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74, 525–556.
- Pintrich, P. R. (2003a). Motivation and classroom learning. In W. M. Reynolds & G. E. Miller (Eds.), *Handbook of psychology: Vol. 7. Educational psychology* (pp. 103–122). New York: Wiley.
- Pintrich, P. R. (2003b). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology*, 95, 667–686.
- Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., et al. (2006). *PISA 2003: Dokumentation der Erhebungsinstrumente [PISA 2003: Scale documentary]*. Münster, Germany: Waxmann.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. London: Sage.
- Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.
- Reeve, J. (1998). Autonomy support as an interpersonal motivating style: Is it teachable? *Contemporary Educational Psychology*, 23, 312–330.
- Reeve, J. (2002). Self-determination theory applied to educational settings. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 183–203). New York: University of Rochester Press.
- Reeve, J., Bolt, E., & Cai, Y. (1999). Autonomy-supportive teachers: How they teach and motivate students. *Journal of Educational Psychology*, 91, 537–548.
- Reeve, J., & Jang, H. (2006). What teachers say and do to support students' autonomy during a learning activity. *Journal of Educational Psychology*, 98, 209–218.
- Reeve, J., Jang, H., Carrell, D., Jeon, S., & Barch, J. (2004). Enhancing students' engagement by increasing teachers' autonomy support. *Motivation and Emotion*, 28, 147–169.
- Reeve, J., Jang, H., Hardre, P., & Omura, M. (2002). Providing a rationale in an autonomy-supportive way as a strategy to motivate others during an uninteresting activity. *Motivation and Emotion*, 26, 183–207.
- Renninger, K. A. (2000). Individual interest and its implications for understanding intrinsic motivation. In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 373–404). San Diego, CA: Academic Press.
- Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology*, 57, 749–761.
- Ryan, R. M., & Deci, E. L. (2000a). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25, 54–67.
- Ryan, R. M., & Deci, E. L. (2000b). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55, 68–78.
- Ryan, R. M., & Grolnick, W. S. (1986). Origins and pawns in the classroom: Self-report and projective assessments of individual differences in children's perceptions. *Journal of Personality and Social Psychology*, 50, 550–558.



- Sansone, C., & Thoman, D. B. (2005). Interest as the missing motivator in self-regulation. *European Psychologist, 10*, 175–186.
- Sansone, C., Weir, C., Harpster, L., & Morgan, C. (1992). Once a boring task always a boring task? Interest as a self-regulatory mechanism. *Journal of Personality and Social Psychology, 63*, 379–390.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*, 147–177.
- Schiefele, U. (1991). Interest, learning, and motivation. *Educational Psychologist, 26*, 299–323.
- Schraw, G., & Lehman, S. (2001). Situational interest: A review of the literature and directions for future research. *Educational Psychology Review, 13*, 23–52.
- Silvia, P. J. (2006). *Exploring the psychology of interest*. New York: Oxford University Press.
- Stefanou, C. R., Perencevich, K. C., DiCintio, M., & Turner, J. C. (2004). Supporting autonomy in the classroom: Ways teachers encourage student decision making and ownership. *Educational Psychologist, 39*, 97–110.
- Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality, 13*, 389–408.
- Stipek, D. (1996). Motivation and instruction. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 85–113). London: Prentice Hall International.
- Stipek, D. (2002). Good instruction is motivating. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 309–332). San Diego, CA: Academic Press.
- Trautwein, U., & Lüdtke, O. (2007). Students' self-reported effort and time on homework in six school subjects: Between-student differences and within-student variation. *Journal of Educational Psychology, 99*, 432–444.
- Trouilloud, D., Sarrazin, P., Bressoux, P., & Bois, J. (2006). Relation between teachers' early expectations and students' later perceived competence in physical education classes: Autonomy-supportive climate as a moderator. *Journal of Educational Psychology, 98*, 75–86.
- Tsai, Y.-M., Kunter, M., Lüdtke, O., & Trautwein, U. (2007, April). *Interest and anxiety in math lessons over three weeks*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.
- Turner, J. C. (2001). Using context to enrich and challenge our understanding of motivational theory. In S. Volet & S. Jarvela (Eds.), *Motivation in learning contexts: Theoretical advances and methodological implications* (pp. 85–104). New York: Pergamon Press.
- Turner, J. C., & Meyer, D. K. (2004). A classroom perspective on the principle of moderate challenge in mathematics. *Journal of Educational Research, 97*, 311–318.
- Turner, J. C., Meyer, D. K., Cox, K. E., Logan, C., DiCintio, M., & Thomas, C. T. (1998). Creating contexts for involvement in mathematics. *Journal of Educational Psychology, 90*, 730–745.
- Urdan, T., & Turner, J. C. (2005). Competence motivation in the classroom. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 297–317). New York: Guilford Publications.
- Vansteenkiste, M., Simons, J., Lens, W., Soenens, B., & Matos, L. (2005). Examining the motivational impact of intrinsic versus extrinsic goal framing and autonomy-supportive versus internally controlling communication style on early adolescents' academic achievement. *Child Development, 76*, 483–501.
- Vermunt, J. D., & Verloop, N. (1999). Congruence and friction between learning and teaching. *Learning and Instruction, 9*, 257–280.
- Wigfield, A., & Eccles, J. S. (Eds.). (2002). *Development of achievement motivation*. San Diego, CA: Academic Press.
- Williams, G. C., & Deci, E. L. (1996). Internalization of biopsychosocial values by medical students: A test of self-determination theory. *Journal of Personality and Social Psychology, 70*, 767–779.

## Appendix

### Items of the Lesson-Specific Measures

#### *Interest Experience (5 items)*

- I enjoyed the topic.
- It was interesting to me.
- The topic was meaningful to me.
- It was important to me that I thoroughly understood the material covered.
- I saw that the content of the lesson can be useful in real life.

#### *Perceived Autonomy-Supportive Climate (6 items)*

- I felt that my teacher provided me choice and options.
- I felt understood by my teacher.
- My teacher conveyed confidence in my ability to do well in the course.
- My teacher encouraged me to ask questions.
- My teacher listened to how I would like to do things.
- My teacher tried to understand how I see things before suggesting a new approach.

#### *Perceived Controlling Behaviors (4 items)*

- Our teacher expected split-second answers.
- Our teacher's instructions were so vague that nobody knew what to do.
- Our teacher covered so much material that we had difficulty keeping up.
- Our teacher was mean to a student.

#### *Perceived Cognitive Autonomy Support (4 items)*

- We worked through exercises that helped us understand the topic.
- Different students presented their solutions to the same task.
- Our teacher set tasks that required time to reflect.
- Our teacher emphasized the relations between the topics discussed.

Received March 29, 2007

Revision received November 13, 2007

Accepted December 16, 2007 ■

# Cognitive Processing About Classroom-Relevant Contexts: Teachers' Attention to and Utilization of Girls' Body Size, Ethnicity, Attractiveness, and Facial Affect

Shirley S. Wang, Teresa A. Treat, and Kelly D. Brownell  
Yale University

This study examines 2 aspects of cognitive processing in person perception—attention and decision making—in classroom-relevant contexts. Teachers completed 2 implicit, performance-based tasks that characterized attention to and utilization of 4 student characteristics of interest: ethnicity, facial affect, body size, and attractiveness. Stimuli were 24 full-body photos of girls that varied along the dimensions of interest. Teachers completed a similarity-ratings task and 4 preference-ratings tasks. Results showed that teachers attended to ethnicity and body size but did not utilize this information when selecting students across contexts. In contrast, teachers relied heavily on affect and attractiveness when making decisions. These results suggest that further investigating cognitive processing in person perception is important in understanding how teachers attend to and use multiple salient student attributes in real-world contexts. This study also illustrates the potential utility of adopting a method that places a premium on internal validity to investigate questions relevant to educational researchers. Future work should investigate how other student information, such as student background or personality, affects teachers' cognitive processing in different classroom-relevant contexts.

*Keywords:* obesity, attractiveness, education, ethnicity, attention

Perception and categorization of people on the basis of salient person attributes is a useful and necessary heuristic for filtering information from complex social stimuli and making sense of the social world (see Fiske, 1993; Macrae & Bodenhausen, 2000, 2001, for a review). McFall's (1982, 1990) social information-processing model provides one way of examining the person-perception process. This framework posits that a series of basic cognitive processes—decoding, decision making, and enactment—operate on the continuous stimulus input received from the environment to produce a behavioral output (i.e., behavior). Understanding the way teachers encode specific student characteristics and utilize those attributes in decision making could better our understanding of teacher behaviors in the classroom. Investigating such basic processes is critical because biased or stigmatizing behavioral output can arise from a combination of biased decoding and decision making based on the input perceived in a specific environment, such as students' salient physical characteristics.

A large body of research on the educational environment has relied on observations of teacher–student interactions in the class-

room (e.g., Blatchford, 2003; O'Connor, Fish, & Yasik, 2004). This work has demonstrated clearly the complexity of teaching and learning in the classroom and has shown that even well-intentioned and subtle teacher behaviors can have unintended negative consequences for students. Our current work builds on such research by attempting to elucidate various cognitive processes that underlie teacher behaviors in the classroom. The many potential facets of student–teacher interactions in real-world situations make it challenging to draw definitive conclusions on the basis of observational methods alone. More standardized and controlled, albeit less externally valid, studies can complement observational research. Studies can help to clarify our interpretation of the observed relationships as well to enhance our understanding of the mechanisms underlying the relationships. Notably, the generalizability of findings from studies that emphasize internal validity can be quite limited, because such studies necessarily control tightly for many important variables to enhance the interpretability of the findings. Thus, one cannot assume that teachers' behavior in experimental settings translates to the real world. The present research employs a compromise between these two research-strategy extremes that integrates strengths of observational and experimental work by systematically examining important teacher, student, and classroom variables in experimental settings that attempt to replicate important aspects of the real-world situation.

The present study examines two aspects of cognitive processing in person perception, attention and decision making, in a contextually meaningful domain—the classroom. Prior to making decisions based on student attributes, teachers must first attend to this information and then choose, consciously or unconsciously, to use that information as a basis for their decisions. We used methodological and analytic paradigms drawn from cognitive science to

---

Shirley S. Wang, Teresa A. Treat, and Kelly D. Brownell, Department of Psychology, Yale University.

This work was supported by the American Psychological Association Graduate Students' Scholarship for Research in Psychology. We thank Richard Viken, Richard Eibach, Mitch Prinstein, and David Armor for their thoughtful comments about this study and Jennifer Barta for her able assistance with data collection. We also acknowledge the students who served as models in the stimulus set in this study.

Correspondence concerning this article should be addressed to Shirley S. Wang, Department of Psychology, Yale University, P.O. Box 208205, New Haven, CT 06520. E-mail: shirley.s.wang@aya.yale.edu



investigate how four student attributes—body size, ethnicity, attractiveness, and facial affect—influenced teachers' perceptions of students and teachers' decision making in four common classroom-relevant contexts. We aimed to complement existing observational research in educational psychology by more thoroughly investigating the processes that underlie teachers' observable classroom behavior. The current work focuses on the following three overarching conceptual questions: (a) On what student attributes do teachers focus when not directed to attend to any specific attribute? (b) Does attention to student-specific attributes necessitate the utilization of the information in decision making? (c) Does utilization of student-specific information vary as a function of the decision-making context? Because teachers must make numerous decisions in the classroom in no more than a few seconds, we decided to focus on the process of rapid decision making in this study. However, teachers engage numerous other complex cognitive processes that also are important to student learning. Our examination of such simplified decision-making scenarios is intended to be only a first step toward better understanding the processes underlying teacher behavior, which ultimately may help inform such activities as training new teachers.

Prior research has focused primarily on the processing and use of physical attributes in isolation from one another, such as investigations of perception of either ethnicity or gender, but not both. Few studies have examined individual differences in attention to multiple important person attributes simultaneously. However, in the real world, perceivers must make sense of competing person attributes, each of which conveys information about the individual. In addition, person perception often is assumed to be a traitlike function of the perceiver (e.g., an individual who scores highly on the Modern Racism Scale is more likely to exhibit discriminatory behavior), and the basis for judgments frequently is assumed to be constant across situations. Yet one could imagine that relative utilization of person attributes depends in part on the context, beyond stable individual differences in the utilization of attributes during decision making. This perspective is consistent with the increasing emphasis in educational research on the situational or contextualized nature of teaching and learning (e.g., Barab, Hay, & Yamagata-Lynch, 2001; Barab & Plucker, 2002; Roth, 1998). Similarly, many salient physical attributes, such as body size and attractiveness, covary naturally in the real world (e.g., thinner individuals are viewed as more attractive), and the relative contribution of each attribute to different stages of cognitive processes and overt behavior has rarely been quantified. Finally, the majority of the existing work on processing of ethnicity, attractiveness, and body-size information has utilized explicit self-report measures rather than implicit performance-based measures (e.g., D. F. Chang & Sue, 2003; Langlois et al., 2000). Our focus in this study on performance-based cognitive methods contributes to person-perception research by examining the relative importance of particular person attributes to decision making in standardized but contextually bound situations. Incorporating cognitive-processing methods into studies of person perception may allow researchers to obtain a richer understanding of the cognitive dynamics of teacher behavior in interactions with students.

## Influence of Student Attractiveness, Body Size, and Ethnicity on Teacher Behavior

In educational contexts, teachers' and administrators' perceptions of student attributes often influence their behavior, either implicitly or explicitly. Most research in this area has focused on teacher expectations as determinants of student academic performance, or the Pygmalion effect, in which students whom teachers expect to perform better do, indeed, perform better (Rosenthal & Jacobson, 1966; see Jussim, 1991, for a review). Expectation formation can occur quickly and with little information (Jussim, 1989). In addition, studies on "thin slices" of behavior show that perceptions and evaluations of individuals can form very quickly from limited information (e.g., Ambady, Hallahan, & Conner, 1999; Ambady & Rosenthal, 1992; Borkenau, Mauer, Riemann, Spinath, & Angleitner, 2004; Friedman, DiMatteo, & Mertz, 1980), even in the classroom. For example, Ambady and Rosenthal (1993) found that judges' evaluations of teachers after examining silent video clips as brief as 2 s correlated with students' end-of-semester ratings of these same teachers. Teachers, similarly, readily form expectations of students on the basis of physical characteristics, such as attractiveness, gender, and ethnicity (e.g., Ritts, Patterson, & Tubbs, 1992; Tauber, 1998).

Ethnicity is a critically important variable to study in educational contexts. Educational attainment and achievement of underrepresented minority groups continue to lag behind those of Caucasians at all socioeconomic levels, despite decades of effort to minimize the discrepancy (Education Trust, 2004; U.S. Department of Education, National Center for Education Statistics, 2001). Numerous researchers have examined how teachers' evaluations and expectations of student aggression, ability, and mental health vary as a function of student ethnicity (e.g., D. F. Chang & Sue, 2003; Graham, 1990; Neal, McCray, Webb-Johnson, & Bridgest, 2003). Yet little is known about the relative importance of ethnicity in comparison to other attributes, such as attractiveness or body size, in the classroom. In addition, research in this area often assumes that attention to ethnicity necessitates that the information will be utilized in decision making, without obtaining independent assessments of attention and utilization. Might it be the case that teachers encode ethnicity information but do not use it to make decisions?

An extensive literature shows that attractive children are perceived to be more popular, intelligent, and friendly than less attractive children (e.g., Langlois et al., 2000), and teachers are no exception to these perceptions (see Ritts et al., 1992, for a review). Facial affect has also been found to play a role in student-teacher interactions. A substantial literature points to the importance of understanding nonverbal communication, such as facial expression and posture, on the part of both teachers and students in their interactions in the classroom (e.g., Goldin-Meadow, 2004; Mottet & Richmond, 2000; Tauber, 1998). For instance, Stuhlman and Pianta (2002) interviewed and observed teachers in the classroom and found that negative affect from students was related to greater negative affect from teachers as well as increased behavioral interactions with the teacher. However, affect and attractiveness covary naturally with each other as well as with other variables, such as body size and perhaps socioeconomic status (via the niceness of clothing or grooming). Teasing apart the relative contributions of these naturally covarying or proxy variables is



important to our understanding of the extent to which attractiveness, versus other attributes, influences teachers' cognitive processing.

Another similarly salient physical characteristic, which is associated with traits such as intelligence and friendliness but which has been relatively neglected in previous research, is that of body size (Hebl & Mannix, 2003; Puhl & Brownell, 2001). The attractiveness literature typically treats body size as a control variable by using only head shots as stimuli or by including only normal-weight children in full-body stimulus sets. As a result, little is known about the extent to which teachers attend to and use body-size information independently of attractiveness information when making classroom decisions. Body size is a stigmatized physical attribute that deserves greater exploration in the literature. Being overweight in childhood is associated with a number of negative outcomes, such as social and academic difficulties (Neumark-Sztainer et al., 2002; see Puhl & Brownell, 2001, for a review). Overweight girls complete fewer years of schooling than their average-weight counterparts (Gortmaker, Must, Perrin, Sobol, & Dietz, 1993) and are less likely to receive help in applying to colleges (Benson, Severs, Tatgenhorst, & Loddengaard, 1980). Obese individuals of both genders are less likely to be accepted to college than their nonobese peers, despite equivalent academic performance (Canning & Mayer, 1966). With the rate of childhood obesity skyrocketing—15% of children now are considered obese (Strauss, 2002)—it is increasingly important to understand how teachers process students' body-size information.

The present study extends previous work by examining teachers' attention to and utilization of student-specific attributes in standardized, performance-based assessments and by comparing teachers' relative utilization of important student characteristics, such as attractiveness, body size, and ethnicity, when making rapid decisions. Our study is the first of which we are aware that attempts to tease apart the influences of body size, affect, and attractiveness on teacher attention and decision making. Although literature exists for many stigmatized attributes, these areas of research rarely have been integrated with one another. In this way, the literature fails to reflect the fact that characteristics of the students are multidimensional. This study explores in a standardized manner the process by which certain attributes are attended to over others and how these attributes influence decision making. The paradigms used to examine these cognitive processes differ from what teachers experience in the classroom, which constrains the external validity of our findings. However, these paradigms serve nicely as experimental analogues of one type of information processing that teachers frequently employ in the educational environment—that is, making rapid decisions, ones in which they do not have extensive real time to deliberate, based on limited student information. Teachers' processing of young female students' body size and ethnicity is of particular interest in this study, given overweight and ethnic-minority students' reported academic and social difficulties, as reviewed previously. We anticipated that teachers would attend to students' body size and ethnicity, rely heavily on this information when selecting students in common classroom situations, and show strong preferences for lighter weight and nonminority students.

### Conceptualization and Measurement of Individual Differences in Attention and Decision Making

Similarity- and preference-rating paradigms, well established in the cognitive literature, are used in this study to characterize teachers' attention and decision-making processes in the classroom. These paradigms are used in conjunction with multidimensional scaling (MDS) methods to evaluate individual differences in teachers' attention to particular characteristics of a set of stimuli (e.g., photos, objects, words, descriptions) as well as variability in their utilization of stimulus attributes when making decisions (Bechtel, Tucker, & Chang, 1971; Davison, 1992; Nosofsky, 1992; Treat, McFall, Viken, & Kruschke, 2001; Treat et al., 2002; Viken, Treat, Nosofsky, McFall, & Palmeri, 2002). These paradigms afford performance-based examinations of attention and decision-making processes, in which researchers draw inferences about the operation of participants' cognitive processes by observing their performance on information-processing tasks that necessitate the input of these processes, rather than relying on participants' verbal reports of the operations or products of such processes. These paradigms allow for the assessment of teachers' attention and decision-making processes in a relatively implicit and indirect fashion, as the task instructions neither specify the stimulus attributes of interest nor direct teachers to attend to or use particular child characteristics. Scaling algorithms can be used to quantify individual differences in the extent to which teachers attend to and use the various unspecified child characteristics when completing the similarity- and preference-rating tasks. Additionally, teachers' strategic presentation of their perceptions and judgments are minimized by the use of relatively brief stimulus presentation times (e.g., Fazio & Dunton, 1997). Thus, these relatively implicit, performance-based assessments presumably provide more valid assessments of teachers' cognitive processing in the classroom than more explicit and direct measures, which focus participants' attention on specific student characteristics of interest (e.g., asking participants to rate the extent to which they attend to or use ethnicity or attractiveness to choose between students).

We developed a photo stimulus set of 6- and 7-year-old girls for use in the similarity- and preference-ratings tasks. The final photo set varied significantly along the four dimensions of primary theoretical interest and minimally along other potential dimensions (e.g., clothing and background characteristics). By design, the students depicted in the photos were unknown to the teachers. Thus, all teachers received identical information about the students, and potentially idiosyncratic background knowledge about the students could not influence teachers' judgments. Much previous educational research has used students or photos of students whom the teachers already knew, which introduces numerous alternative explanations for any findings. The use of a standardized stimulus set, rather than a naturally occurring one, also allows the investigator to ensure that there is sufficient variation along each of the dimensions of interest. Our approach allowed us to separate variables that might covary naturally, such as attractiveness and either body size or facial expression, as well as to adequately cover the entire multidimensional space of interest, which facilitated clearer interpretation of the results. Thus, by using a standardized stimulus set and a controlled experimental environment, we could gain a better understanding of the extent to which the teacher calls on Susie Student because of Susie's body size, rather than her attrac-



tiveness or facial expression. Teasing apart the extent to which these naturally covarying attributes affect teachers' perceptions and behaviors supplements existing quasi-experimental research and may shed additional light on our understanding of teachers' decision making in the classroom.

The attributes of primary interest—body size, ethnicity, and attractiveness—were built into the stimulus set. Facial affect was also incorporated because it is a potential indicator of student interest and enthusiasm that is relevant to teachers' decisions (e.g., see Mottet & Richmond, 2000, for a review). It was not possible to represent attractiveness completely independently of body size, ethnicity, and affect in the present stimulus set. Nonetheless, the significant nonoverlap between the attributes raises the possibility that results for attractiveness might diverge from those for the other dimensions. We recognize that teachers likely attend to and use many other child-specific characteristics, such as the child's personality and past behavior or performance. Because the focus of the present study is on understanding how teachers cognitively process students' observable physical characteristics, however, we wanted to rule out competing explanations for teachers processing by eliminating from the stimuli other potential influences on their processing. Other attributes of particular interest could be incorporated readily into the stimulus set in a standardized fashion in future research.

Evaluating Individual Differences in Attention to Stimulus Dimensions

Participants' representation of and attention to stimulus dimensions can be examined within a similarity-ratings paradigm, in

which participants rate the similarity of stimulus pairs on a 10-point scale, where 1 = *not at all similar* and 10 = *extremely similar* (Davison, 1992; Schiffman, Reynolds, & Young, 1981). Participants' similarity ratings provide an indirect indicator of their relative attention to the stimulus dimensions. For example, a participant who judges two happy-looking girls of different ethnicities to be quite similar likely is attending more to affect than to ethnicity. In contrast, a participant who evaluates an attractive pair of heavier and lighter girls as very dissimilar likely is attending more to body size than to attractiveness.

Participants' similarity ratings served as the input for an MDS analysis, which provided a group-level representation of the stimuli or "psychological space" (Davison, 1992; Nosofsky, 1992; Schiffman et al., 1981; Treat et al., 2002). This multidimensional depiction of the stimulus set was spatial in nature: Stimuli that were perceived to be more dissimilar were scaled farther apart than stimuli that were judged to be more similar. Figure 1A depicts a two-dimensional (2-D) psychological space in which each point corresponds to a unique stimulus (i.e., a photo of a particular girl). Stimuli are plotted along the two dimensions spanning the psychological space: body size and affect. This solution indicates that Girl A and Girl B were viewed similarly, whereas both girls were viewed as quite dissimilar to Girl C.

The weighted MDS approach (WMDS; Carroll & Chang, 1970; Carroll & Wish, 1974) simultaneously represents both the group-level psychological space and individual-specific differences in attention to the stimulus dimensions in the group space. Thus, WMDS estimates not only the group-level multidimensional stimulus coordinates but also individual-specific dimensional weights

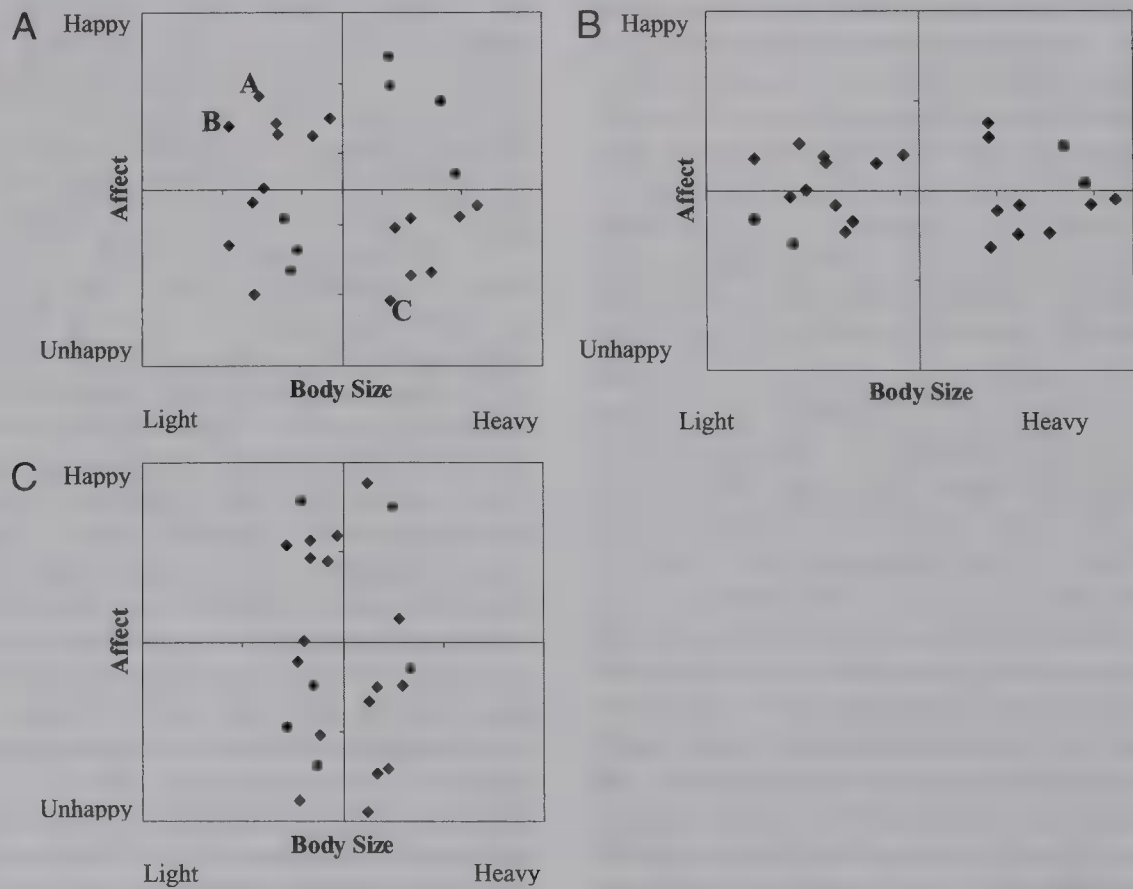


Figure 1. Group- and individual-level scaling representations of psychological space. Panel A: group-level psychological space for 24 photo stimuli. Panel B: individual-level psychological space for the body-size-oriented individual. Panel C: individual-level psychological space for the affect-oriented individual.

that indicate how much an individual attends to each dimension in the group solution. These dimensional weights are applied to the group-level stimulus coordinates and are assumed to stretch or shrink the dimensions of the group-level psychological space. For example, Figure 1B depicts the individual-specific psychological space of a "body-size-oriented" participant who attends relatively more to body size than to affect, in comparison to the average participant. This orientation is characterized by the stretching of the body size dimension and the shrinking of the affect dimension in the shared psychological space in Figure 1A. Note that this representation scales heavy and light girls further apart than happy and sad girls. Thus, the stretching and shrinking of these two stimulus dimensions represent the participant's perception that heavier and lighter girls are relatively more dissimilar than happier and sadder girls, in comparison to the perceptions of the participant group as a whole. In contrast, Figure 1C portrays an affect-oriented participant who views heavy and light girls as more similar than happy and sad girls.

The present study uses the similarity-ratings paradigm to assess teachers' group-level psychological representation of students as well as individual differences in teachers' attention to the student-specific characteristics of body size, ethnicity, affect, and attractiveness. We anticipate that teachers will attend strongly to students' body size and ethnicity.

### Evaluating Individual Differences in Utilization of and Preferred Directions Along Stimulus Dimensions When Making Rapid Decisions

In addition to exploring teachers' attentional processes, we also seek to understand how teachers utilize information about students when making relatively rapid decisions (i.e., in no more than a few seconds). Teachers often make rapid decisions in the classroom in which they have minimal opportunity to deliberate and limited direction as to the basis for these decisions, such as when selecting a child to answer a question in class. Of course, there are many important classroom situations in which teachers have ample time to deliberate in decision making, such as when awarding student grades. Nonetheless, the constraints under which teachers operate in the classroom necessitate that decisions such as those explored in the present study frequently must be made very quickly. Under these circumstances, teachers do not have time to evaluate consciously and deliberately which student attributes will be the focus

of their attention or will be weighted heavily in their decisions. Thus, we limited stimulus presentation times and urged participants to respond quickly in the present study, in an effort to provide an experimental context that is analogous to that encountered by teachers in the classroom. The pressure to respond quickly also minimizes the influence of social desirability and other presentation biases that come into play as stimulus presentation and decision-making times increase.

Attention and decision making are distinct cognitive processes, as teachers might attend to particular student characteristics but not necessarily use them as a basis for making decisions in the classroom. Given that discrimination against overweight and ethnic minority individuals occurs, however, it appears that many people do use body size and ethnicity as a basis for their decisions. Thus, using the same photo set described previously, we investigate the extent to which teachers use student body size, ethnicity, attractiveness, and affect when making rapid decisions in specific classroom contexts.

The preference-ratings paradigm, in conjunction with the PREFMAP scaling program (Carroll, 1972; J. J. Chang & Carroll, 1972; Meulman, Heiser, & Carroll, 1986), frequently is used to assess individual differences in participants' utilization of the dimensions of a previously specified group-level psychological space; the group-level space typically is obtained from a separate MDS analysis of participants' similarity ratings, as described above. Participants again view stimulus pairs in a preference-ratings task, but they select one of the stimuli rather than rate the similarity of the two stimuli. For example, in one of the preference-ratings tasks in the present study, teachers indicated which of two girls they would call on in class. The summary data that were submitted for PREFMAP analysis consisted of the number of times that each teacher chose each girl.

PREFMAP also calculates individual differences in the preferred direction of the previously specified group-level psychological space. PREFMAP computes individual-specific vectors, which specify the direction that is preferred by an individual in the psychological space. Figure 2A displays sample graphic PREFMAP output for one teacher. Each point again corresponds to a unique stimulus, and the overall stimulus configuration is assumed to be the same group-level psychological space that was uncovered in the similarity-ratings solution depicted in Figure 1A. A teacher's preference data are represented by the arrow that points toward the

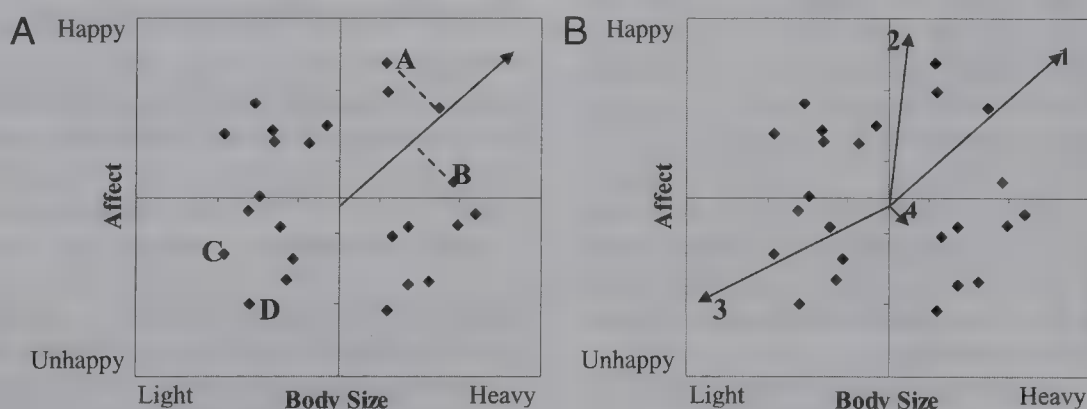


Figure 2. Vector model representation of teachers' preferences. Panel A: group-level psychological space and preference vectors for one teacher. Panel B: preference vectors for four teachers.



region of the space that contains the girls she is more likely to choose and away from the region of the space that contains the girls she is less likely to choose. According to the PREFMAP vector model, stimuli that exhibit perpendicular projections to the vector that are closer to the preferred direction are selected with greater frequency than stimuli with perpendicular projections that are farther from the preferred direction. Thus, the teacher in Figure 2A selected heavy, happy girls (e.g., A and B) more frequently and light, unhappy girls (e.g., C and D) less frequently.

We analyze both the signed and the unsigned values of the vector heads in the present study, as they provide different types of information. The absolute values of the vector endpoints indicate individual differences in the utilization of the stimulus dimensions. Figure 2B presents four teachers, each of whom is represented as a normalized vector in this space, and indicates that both Teachers 1 and 3 relied heavily on both body size and affect when making decisions, because their vector endpoints are extreme along both the body size and the affect dimensions. In contrast, Teacher 2 relied almost exclusively on affect when making decisions: The vector endpoint is extreme along the affect dimension and near zero along the body size dimension. Finally, Teacher 4 used neither body size nor affect in her decision making, as the vector endpoints are near zero along both dimensions. Thus, the unsigned vector endpoints for a particular teacher tell us about the teacher's utilization of the stimulus dimensions when making decisions. In contrast, the signed vector endpoints along each dimension indicate individual differences in teachers' preferred directions along the stimulus dimensions. In the present example, focusing on the signed endpoints allows us to distinguish Teachers 1 and 3, who both relied heavily on body size and affect when making decisions but who selected heavy, happy and light, unhappy girls, respectively.

The preference-ratings paradigm is used in the current study to examine teachers' use of girls' body size, ethnicity, affect, and attractiveness when making rapid decisions in four classroom-relevant contexts. Context selection was informed by the literature on stereotyping based on ethnicity, attractiveness, and body size (Puhl & Brownell, 2001; Ritts et al., 1992; Zebrowitz, Hall, Murphy, & Rhodes, 2002) as well as informal discussions with numerous teachers about the kinds of rapid decisions they make routinely in the classroom. We designed and selected contexts in which it seemed likely that teachers would use the dimensions of interest to us in this study. Teachers in this study also were asked to imagine that they were substitutes who did not know these students, as this provided a plausible rationale for making decisions about students with whom they were unfamiliar. We recognize that there are numerous other classroom-relevant contexts and teacher roles that we could have examined. Given our focus on understanding the underlying processes of how teachers use specific student information to inform their decisions, however, we chose to design a tightly controlled study, despite its limited generalizability. Future research can and should examine teachers' perceptions and behaviors in other contexts involving different cognitive processes, such as situations that call for more deliberate and time-consuming decision making.

In the current work, one context required a teacher decision that conveyed beliefs about students' perceived abilities (i.e., selecting a student to complete a difficult problem on the board). Another required a decision that communicated teacher perceptions of the

student's responsibility or dependability (i.e., asking a student to take a note down the hall). A third context required a teacher decision that potentially resulted in special attention or encouragement for a student (i.e., calling on a student to answer a question). The final context targeted a teacher decision that conveyed who the teacher felt would like or deserve special attention (i.e., assigning the lead role in a class skit), which we chose specifically because we hypothesized that physical appearance would be a factor in the teachers' decision making.

## Present Study

The present study aimed to characterize three aspects of teachers' cognitive processing in the classroom that have not been explored previously: (a) teachers' relative attention to young girls' body size, ethnicity, attractiveness, and facial affect; (b) teachers' relative utilization of these characteristics when selecting girls in four commonly occurring classroom contexts; and (c) teachers' preferred directions along these dimensions when selecting girls in the four contexts. Additionally, we explored the link between individual differences in these cognitive variables and teachers' self-reported teaching experiences and demographic information. Given the marked social and academic difficulties reported by heavier and ethnic minority students, we anticipated that teachers would (a) attend strongly to body-size and ethnicity information in the similarity-rating task; (b) utilize body-size and ethnicity information heavily when selecting children in the preference-rating tasks; and (c) exhibit a preference for lighter, nonminority children in the preference-rating tasks.

## Method

### Participants

Twenty-seven elementary school teachers (21 women, 6 men) from nine different schools within a local school district participated in the study; they were predominantly Caucasian (24 Caucasians, 1 Latina, and 2 Asian Americans). Their mean age was 36.8 years ( $SD = 12.1$ ), and they reported an average of 10.2 years ( $SD = 9.2$ ) of teaching experience. Participants were informed of the study at their monthly staff meeting and by subsequent individual contact at lunch or after school. They were invited to participate in the study during nonschool hours in exchange for \$20 or a book donation to their classroom.

### Development of Experimental Stimuli

Twenty-six 6- and 7-year-old female models were recruited from an elementary school outside the district in which the study was conducted for development of the stimulus set. Written parental permission and verbal child assent were obtained for each child. Each model was photographed individually with standard lighting while standing and facing the camera at the same fixed distance. All models were asked to wear plain white T-shirts, blue jeans, and light-colored sneakers to minimize variability along dimensions that were irrelevant to the study. Models were asked to vary their facial expression across photographs (e.g., look happy, look sad, and look "neutral"). The stimulus set was quite ethnically diverse; 10 of the 26 models were African American or Asian American. Multiple photos of each child were taken, which pro-

vided us with a large number of photos from which to select a stimulus set that varied sufficiently along the dimensions of interest.

To enhance the variability along the body-size dimension, we digitally altered the photos using the WinMorph program (Kumar, 2002) to increase and decrease the original body size of the models. Only body size was warped; faces and facial expressions remained untouched. Each model was warped to appear both heavier and lighter than her current body size so that there were nine possible combinations of body size and affect for each model (happy, sad, and neutral affect crossed with actual, thinner, or heavier body size).

*Normative data collection.* Normative data were collected to quantify the stimulus values along each dimension and to ensure that the dimensions of interest in the stimulus set varied sufficiently and as independently as possible. Data were collected on those photos that did not appear grossly unrealistic or inappropriate to the experimenter (e.g., facial expression was incongruent with the desired expression).

Twenty-seven undergraduate women rated 140 photos along four 10-point scales for the following attributes: body size (1 = *light*, 10 = *heavy*), affect (1 = *unhappy*, 10 = *happy*), friendliness (1 = *unfriendly*, 10 = *friendly*), and attractiveness (1 = *unattractive*, 10 = *attractive*). Participants also rated the perceived ethnicity of the models using any possible combination of five racial groups (Asian, Hispanic, African American, Caucasian, and Native American), the age of the model (4 to 13 years old), and whether the photo looked realistic (i.e., yes or no). Photos were presented in a random order. Participants were instructed to respond as quickly as possible, because we were interested only in their first impressions, and they were told that there were no right or wrong answers.

*Stimulus set selection.* The final stimulus set included 24 photos of unique girls. Stimulus selection was based on the normative data and subject to several constraints. Stimuli were selected to cover the entire 2-D Body Size  $\times$  Affect space as completely as possible.<sup>1</sup> Girls of different ethnicities were distributed proportionally across this space. The judged realism of the stimuli also was taken into account in the selection process; we eliminated those photos that raters deemed unrealistic.

Stimuli also were selected such that the dimensions of primary interest varied as independently as possible. Affect and body size were uncorrelated, as intended, Pearson<sup>2</sup>  $r(24) = -.004$ . Friendliness and affect correlated nearly perfectly,  $r(24) = .988$ ,  $p < .01$ ; we focused only on the affect dimension in pilot testing and actual testing procedures. Attractiveness correlated with both affect and body size,  $r(24) = .599$ , and  $r(24) = -.421$ , respectively, although these correlations were minimized to the extent possible during the stimulus selection process. Age did not correlate significantly with any of the attributes.

## Cognitive Tasks

*Similarity-ratings task.* Participants viewed multiple pairs of stimuli on a laptop computer and rated their similarity on a 10-point scale (1 = *not at all similar*, 10 = *extremely similar*). Participants viewed a random subset of 100 of the 276—that is,  $(24 \times 23)/2$ —pairwise stimulus combinations. They were asked to respond as quickly as possible and were told that we were inter-

ested only in their first impressions and that there were no right or wrong answers. Each stimulus pair was also presented for only five seconds to encourage participants to respond more rapidly and to avoid lengthy deliberation about their responses.

Eight undergraduate students completed a pilot version of the similarity-ratings task for credit in their introductory psychology class. This pilot work ensured the following: (a) Ethnicity, body size, and affect emerged as dimensions in the psychological space; (b) the stimulus set adequately covered the psychological space; (c) the dimensions varied as independently as possible; and (d) the MDS solution based on the first 100 trials was nearly identical to the solution yielded from all trials.<sup>3</sup>

*Preference-rating task.* Participants were asked to assume that they were substitute teachers for the day and had never seen this group of students (i.e., the stimulus set) before. Participants viewed all 276 possible pairs of stimuli for 2 s apiece<sup>4</sup> and, for each pair, selected one of the two displayed stimuli. The presentation time was established on the basis of previous research and practical considerations. Social perception research has found that behavioral snapshots as brief as 2 s yield meaningful information for making evaluations (e.g., Ambady & Rosenthal, 1993; Friedman et al., 1980) and that longer observations do not lead to more accurate prediction of social outcomes (see Ambady & Rosenthal, 1992, for a review). Information processing related to person perception also often uses photo stimuli, and these stimuli are commonly presented for periods of time similar to the duration in which the stimuli in the current study were presented (e.g., Cloutier, Mason, & Macrae, 2005; Macrae, Quinn, Mason, & Quadflieg, 2005; Michel, Caldara, & Rossion, 2006). Given the number of trials that participants would have to complete, it also was important to select the briefest significant presentation time possible to decrease participant fatigue while ascertaining meaningful data. Thus, using the social perception literature and practical considerations as guides, we selected 2 s as the presentation duration in this task. This relatively brief presentation time also maximized similarity to the short time scale on which teachers necessarily make such decisions in classroom contexts.

Four independent decision-making contexts were presented to participants in a counterbalanced order:

1. You ask a question in class. These two students raise their hands at the exact same time. They both speak a

<sup>1</sup> The affect dimension was adequately represented by girls with neutral and happy expressions, as stimuli with neutral expressions were perceived as very unhappy. Thus, it was unnecessary to include stimuli with sad expressions in the stimulus set.

<sup>2</sup> All correlations reported in the text are Pearson correlations, unless otherwise indicated.

<sup>3</sup> Pilot data showed that the MDS solutions based on 100 trials were nearly identical to those estimated from the full 276 trials. Thus, participants in the primary study were asked to rate only 100 pairs in the interest of minimizing participant fatigue.

<sup>4</sup> We conducted pilot work to examine whether it was feasible for participants to complete the ratings within this short time frame. The data analyses yielded readily interpretable and adequately fitting multidimensional solutions, indicating that participants were able to complete the task within this time frame.



similar amount in class. Who do you call on to give the answer?

2. You need a student to take a note to the principal's office all the way on the other side of the building. Who do you ask?
3. Two students are going to the board to complete a problem in front of the class. One problem is easier and one is hard. Who do you assign to do the hard problem?
4. Each reading group in the class acts out a short story for the others on Fridays. The roles are assigned two days prior, and today is "casting day." Each student is assigned a role, but the "leading lady" is assigned first. Who do you choose to play this part?

Twenty-four undergraduate students completed a pilot version of the four preference-rating tasks for credit in their introductory psychology class. This pilot work demonstrated that the task was feasible under speeded conditions and that participants utilized the stimulus dimensions when making decisions in the four contexts.

### *Self-Report Measures*

All participants completed a questionnaire that assessed demographic information, teaching-relevant information, and their evaluation of the cognitive tasks. Teaching information included years of teaching experience, amount of experience with a diverse student body (e.g., gender, ethnicity, socioeconomic status, and body size), current satisfaction with teaching experience (1 = *not at all satisfied*, 7 = *completely satisfied*), and intention to continue teaching (1 = *do not intend to continue teaching*, 7 = *definitely intend to continue teaching*).

The cognitive-task evaluation consisted of three components: (a) a written description of the student characteristics that teachers used to make decisions in each of the four decision-making contexts; (b) a written description of what additional information they would use in the real world when making similar types of decisions; and (c) a checklist of child characteristics that the teacher deliberately attempted not to use in the preference-rating tasks, including hair color, body weight, height, style of clothing, ethnicity, facial expression, body posture, and hair style.

### *Procedure*

Participants provided informed consent and then completed the similarity-ratings task and the four preference-ratings tasks on a laptop computer. Instructions urged participants to complete the tasks as quickly as possible, as we were interested in their first impressions and there were no right or wrong answers to the questions. We purposely did not direct participants to attend to any of the stimulus dimensions of interest. Finally, participants completed the self-report measures. They were debriefed and paid after completing the experimental tasks.

### *Data Screening*

All data were checked for missing values and possible outliers. Although the distributions of several demographic variables were

skewed, there were no obvious outliers, and thus all data were retained for analysis. Nonparametric analyses were conducted with nonnormally distributed variables. Normality of proximity indexes is not assumed in MDS analyses; therefore, all similarity-ratings and preference data were retained.

## **Results**

### *WMDS Analysis of Dimensional Attention*

*Analysis specifications.* Participants rated the similarity of 100 randomly presented pairs of photos out of a total of 276 possible photo pairs. WMDS was used to characterize the group-level psychological space as well as individual differences in teachers' attention to stimulus dimensions. Similarity ratings were reverse coded prior to analysis (i.e., 10 = 1, 9 = 2, etc.), because WMDS analyzes dissimilarity ratings (Carroll & Chang, 1970; Carroll & Wish, 1974). After recoding, a dissimilarity matrix of the input data was created for each participant; missing data (i.e., 176 ratings per participant) were ignored. Equal dissimilarity values, or "ties" in the data, were "untied," or not constrained to be equal in the analysis. Nonmetric MDS analyses of the 27 resulting matrices estimated both group-level stimulus coordinates and individual-specific dimension weights in two, three, four, and five dimensions.<sup>5</sup> All participants' data were retained, as fit indexes were adequate for all individual participants across all four solutions.

*Selecting solution dimensionality.* We determined the dimensionality of the selected solution by examining the relative fit of the models, the interpretability of the estimated dimensions, and the consistency of the obtained dimensions with the dimensions of theoretical interest. The primary indicator of model was the stress index, a badness-of-fit index that ranges between 0 and 1. Stress quantifies the magnitude of the deviation of the scaled distances from an ordinal transformation of participants' similarity ratings (Davison, 1992; Schiffman et al., 1981). Stress values for the 2-, 3-, 4-, and 5-D solutions were .192, .133, .104, and .088, respectively.<sup>6</sup> The magnitude of the fit indexes was comparable to that reported in other scaling studies of complex social perception (e.g., Treat et al., 2001, 2002). The relative improvement associated with an increase in dimensionality was greatest in the move from a 2-D to a 3-D solution (.059), but modest improvement also was noted for the 4-D solution relative to the 3-D solution (.029). However, both visual inspection of the solution and the results of the correlational analyses described below suggested the superiority of the 4-D solution, which accounted for 85.9% of the variability in participants' similarity ratings.

<sup>5</sup> Nonmetric MDS assumes only an ordinal relationship, rather than a linear relationship, between participants' proximity ratings for stimulus pairs and the scaled distances between stimuli in the MDS solution. Thus, only the rank ordering of participants' similarity ratings is assumed to be meaningful.

<sup>6</sup> The fit of the WMDS model varies not only as a function of the number of retained dimensions but also as a function of the number of stimuli and the number of participants. Moreover, model fit is expected to be worse when perceptually and conceptually complex stimuli are being evaluated, as in the present study. Thus, fit adequacy is best evaluated on a study-by-study basis by ascertaining whether the obtained stress values are decidedly lower than those observed when comparable random data matrices are evaluated (see Treat et al., 2001, for an illustration of this).

*Interpreting solution dimensions.* To characterize the dimensions of the selected 4-D solution, we correlated the 4-D stimulus coordinates with the normative data for body size, affect, and attractiveness. Affect correlated most strongly with the fourth dimension (Pearson  $r = -.59$ ), while body size correlated most strongly with the third dimension ( $r = .91$ ). Attractiveness showed moderate associations with both the first and the third dimensions ( $r_s = -.47$  and  $-.41$ , respectively).

Visual inspection of the psychological space suggested that the first two dimensions in the WMDS solution represented variation in the girls' ethnicity, as girls with similar ethnicities were scaled more closely to one another than girls with different ethnicities. As illustrated in Figure 3, this 2-D representation of the girls' ethnicities was more interpretable as four clusters than as two continuous dimensions. The four clusters contained the 5 African American girls, the 4 Asian American girls, the 5 dark-haired Caucasian girls, and the 10 light-haired Caucasian girls.

In summary, our analyses suggest that the first two dimensions in the WMDS solution represented variability in ethnicity, whereas the third and fourth dimensions represented variation in body size and affect, respectively. Attractiveness clearly was relevant to participants' similarity ratings but did not emerge as a separate fifth dimension in the WMDS analyses. Thus, we did not evaluate individual differences in attention to attractiveness.

*Individual differences in dimensional attention.* We depicted individual differences in attention by allowing dimensions to be weighted between 0.0 and 1.0 for each individual, such that a higher weight indicated greater attention to that dimension. Figure

4 depicts the average dimensional attention weights for the 27 teachers in two 2-D scatterplots. The average dimensional attention weights were similar for the four dimensions: .49, .45, .44, and .41 for Ethnicity 1, Ethnicity 2, body size, and affect, respectively. The similarity of the magnitude of these attention weights indicated that the average participant attended at a similar level to all four dimensions (i.e., the perceived salience of the four dimensions represented in this stimulus set was roughly equivalent). Substantial individual variation in participants' dimensional attention also emerged, particularly for the Ethnicity 1 and body-size dimensions, as indicated by greater standard deviations in the weights for these two dimensions ( $SDs = 0.15$  and  $0.16$ , respectively) than for the Ethnicity 2 and affect dimensions ( $SDs = 0.08$  and  $0.10$ , respectively). Overall, therefore, participants on average clearly were attending to the girls' body size, ethnicity, and affect.

Individual differences in dimensional attentional patterns were unrelated to teachers' age, years of teaching experience, job satisfaction, or amount of experience in working with students of diverse backgrounds. Nonparametric (Spearman's) correlations between attentional subject weights and individual-differences variables were uniformly small and no more frequently significant than would be expected by chance alone. For example, teachers who reported having a large percentage of African American or Asian American students in their classrooms in the last 5 years did not attend more or less to ethnicity than teachers who reported little experience with ethnic minorities. Moreover, teachers who reported avoiding the use of certain student characteristics did not differ in their attention to dimensions compared to teachers who

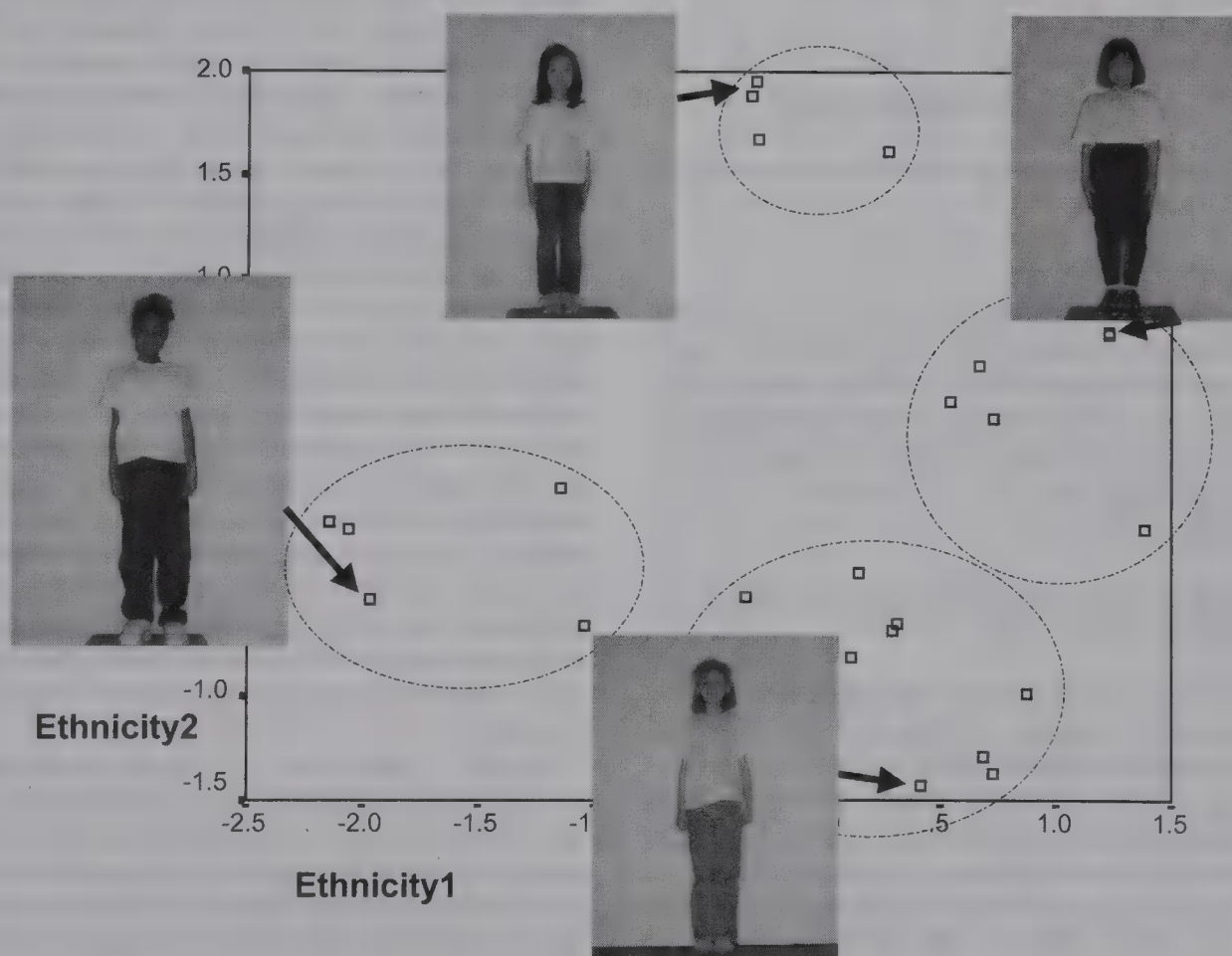


Figure 3. Two-dimensional scaling representation of ethnicity.



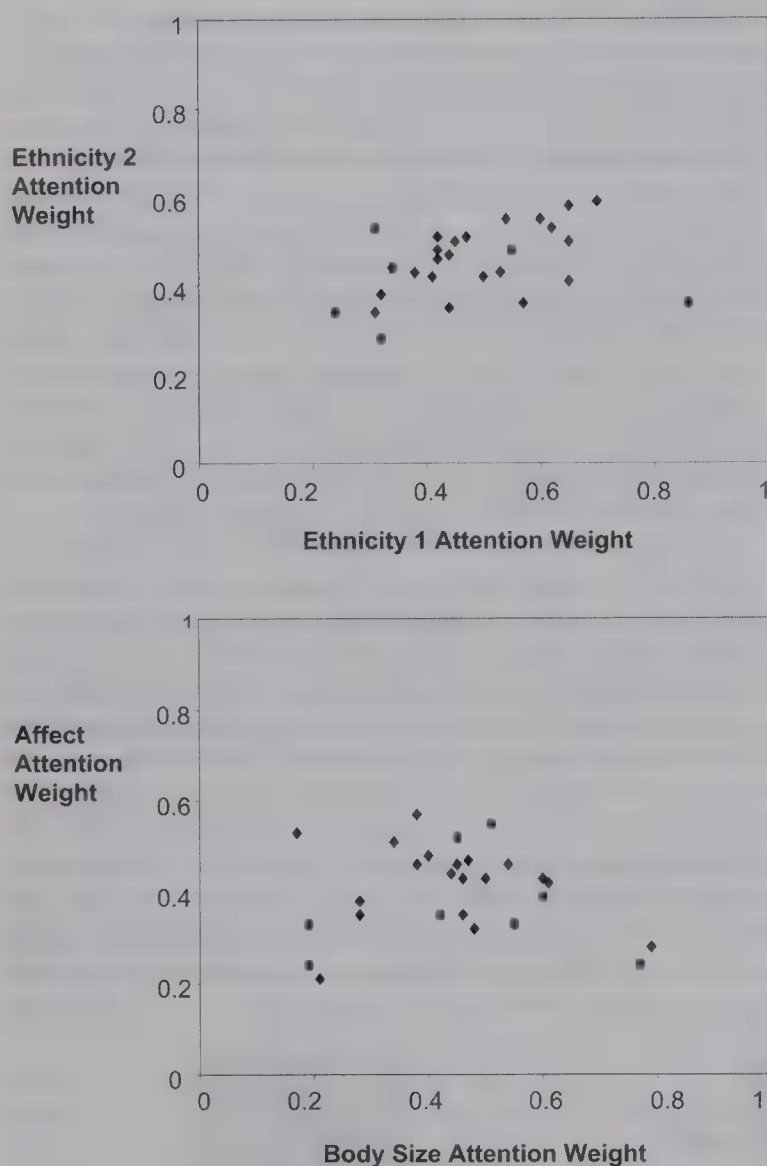


Figure 4. Dimensional attention weights for four-dimensional weighted multidimensional scaling solution.

did not report avoiding the use of characteristics. For instance, the 16 teachers who reported avoiding body size did not attend less to body size than the 10 teachers who did not report avoiding body size. These results suggest that there were substantial individual differences in attention to the student attributes of interest—body size, affect, and ethnicity—that were not associated with basic teacher demographic information.

#### *PREFMAP Analysis of Dimensional Utilization and Preferred Dimensional Directions*

Participants viewed all 276 pairwise combinations of the 24 girls and selected 1 of each pair of girls in each of the four decision-making contexts. The PREFMAP-3 program (Meulman et al., 1986) was used to quantify individual differences in the utilization of and preferred directions along the stimulus dimensions in the four decision-making contexts. To prepare the data for analysis, we totalled the number of times each girl was selected separately for each teacher and context; the resulting values ranged from 0 to 23 selections for the 24 girls in each of the four contexts. Then we created a Participant  $\times$  Stimulus ( $27 \times 24$ ) matrix of

these counts for each of the four contexts. These matrices served as the input data for four separate PREFMAP analyses.

For each of the four contexts, we conducted a 5-D, external, nonmetric PREFMAP analysis (Meulman et al., 1986). In external analyses, the stimulus coordinates are constrained to be equal to relevant values external to the choice data. The “external” approach to preference analysis eliminates the need to estimate both stimulus coordinates and vector endpoints in a single analysis, as the stimulus coordinate estimates are fixed, rather than free, parameter estimates. The resulting reduction in the number of estimated parameters reduces the likelihood of problematic solutions, such as multiple solutions that fit equally well but provide markedly different parameter estimates (Borg & Groenen, 1997; Davison, 1992; Schiffman et al., 1981). Such problems commonly arise when “internal” approaches to preference analysis are used, because so many parameters are being estimated (i.e., the model is far too complex to fit to the available data). Thus, in the present case, only 27 vector endpoints were estimated (i.e., five parameters were estimated per teacher). Stimulus coordinates were derived from two external sources in our analyses: (a) the WMDS solution, which provided stimulus coordinates for body size, affect, and the two dimensions of ethnicity, and (b) the standardized normative ratings for attractiveness. Attractiveness was included in the PREFMAP analyses to facilitate evaluation of teachers’ relative utilization of body size and attractiveness. Ties in the preference data were untied in the analyses.

In the nonmetric approach to preference scaling, PREFMAP-3 selects parameter estimates that maximize the correlation between an ordinal (rather than linear) transformation of participants’ preference data and the model-predicted preferences (Meulman et al., 1986). This goodness-of-fit index was highest for the first context ( $r = .829$ ) and was similar for the remaining three contexts ( $r_s = .755, .750, \text{ and } .788$ , respectively). These values indicated that the five dimensions accounted for a substantial proportion of the variability in participants’ choice data across the four decision-making contexts. Here again, the fit of the model to randomly generated data was markedly worse.

We normalized teacher-specific vectors within teacher to be of unit length across the five imposed dimensions prior to conducting further analyses (i.e., the square root of the sum of the squared endpoint values equaled 1.0). Thus, the absolute values of the five vector endpoints ranged from a minimum of 0.0, which indicated no utilization of a dimension, to 1.0, which indicated sole utilization of a dimension. Normalization of vector length maximized the comparability of the vector endpoint values across teachers, as the length of unnormalized vectors in the PREFMAP model varied as a function of the fit of the model to each teacher’s choice data. Parallel analyses of the unnormalized vector data resulted in identical conclusions, so we present the findings based on the normalized vector data in the following section to maximize their interpretability.

*Analysis of dimensional utilization.* The unsigned values of the vector endpoints indicated teachers’ dimensional utilization of the five dimensions but not the preferred direction along the dimension. For example, these values indicated whether body size was important to a teacher’s decisions but not whether the teacher typically tended to select heavier or lighter girls.

Figure 5 presents average utilization values for the five attributes across the four decision-making contexts. A 5 (attribute)  $\times$

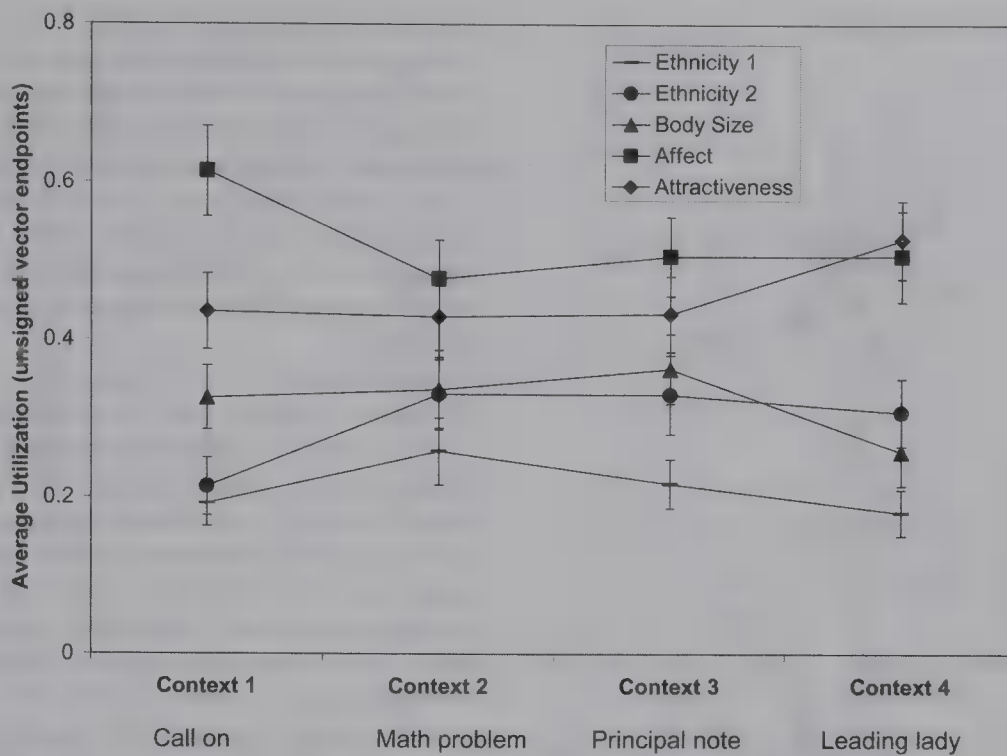


Figure 5. Average dimensional utilization of five attributes in four contexts. Vertical rules depict standard errors of the means.

4 (context) repeated-measures ANOVA indicated a main effect of attribute,  $F(4, 104) = 16.536, p < .001$ , on dimensional utilization. Neither the context effect nor the Attribute  $\times$  Context interaction was significant. Post hoc examination of the attribute effect, using Sidak-based multiple comparisons, indicated that (a) teachers used affect and attractiveness (utilization  $M_s = 0.526$  and  $0.457$ , respectively) significantly more across contexts than the body size, Ethnicity 1, or Ethnicity 2 dimensions ( $M_s = 0.321, 0.296$ , and  $0.212$ , respectively;  $p_s < .05$ ) and (b) teachers used body size significantly more than the Ethnicity 1 dimension ( $p < .05$ ). Minimal individual differences were present in teachers' utilization of the Ethnicity 1, Ethnicity 2, and body size dimensions ( $SD_s = 0.110, 0.113$ , and  $0.122$ , respectively), whereas variability in utilization of affect and attractiveness was greater ( $SD_s = 0.193$  and  $.179$ , respectively) and potentially predictable in later analyses.

Figure 6 presents the PREFMAP representation of teachers' decision making in a representative context, the "principal note" context. The 27 vectors specify the 27 teachers' preferred directions along each of the five dimensions. The figure illustrates teachers' greater utilization of affect and attractiveness (bottom panel), relative to the remaining three attributes, when selecting a student to take a note to the principal's office. In other words, the length of the typical teacher's vector was substantially longer along the affect and attractiveness dimensions than along the remaining three dimensions.

**Analysis of preferred dimensional directions.** The signed values of teachers' vector endpoints indicated their preferred direction along the five dimensions (e.g., whether a teacher tended to choose happy or sad girls). Figure 7 displays the average signed values along the five dimensions for the four decision-making contexts. A 5 (attribute)  $\times$  4 (context) repeated-measures ANOVA indicated effects of both attribute,  $F(4, 104) = 6.472, p < .001$ , and context,

$F(3, 78) = 3.000, p < .05$ . The Attribute  $\times$  Context interaction was not significant. Sidak post hoc evaluations of the attribute effect revealed that the preferred direction along the affect dimension was significantly more positive (i.e., happy) than the preferred directions along body-size and both ethnicity dimensions (all  $p_s < .05$ ) and that the preferred direction along the attractiveness dimension was significantly more positive (i.e., attractive) than the preferred directions along the Ethnicity 1 and body-size dimensions (both  $p_s < .05$ ). In other words, happy, attractive children were more likely to be selected across the four contexts than children of a particular body size or ethnicity. These findings are readily discerned from a reinspection of teachers' preferred directions along the five dimensions in Figure 6. That is, the majority of the vector heads point toward the positive ends along the affect and attractiveness dimensions; few vectors point toward the sad-unattractive quadrant of the space. Post hoc evaluations of the significant context effect indicated that the preferred direction was significantly more positive across dimensions in the first context than in the second context ( $p < .05$ ); this finding simply reflects the greater consensus of teachers' preferred directions along the five dimensions in the first decision-making context (i.e., teachers were much more likely to select attractive, happy girls in the first context).

We supplemented this quantitative analysis of teachers' preferred directions along the five stimulus dimensions with a more qualitative characterization of the data for two reasons. First, the preferred direction along a particular dimension only matters substantively to the extent that a teacher utilizes this dimension significantly when making decisions. In other words, showing a slight preference for lighter children when one's decisions are determined almost entirely by children's affect or attractiveness is of limited concern in the present context. Second, marked disagreement among teachers about the preferred direction along a



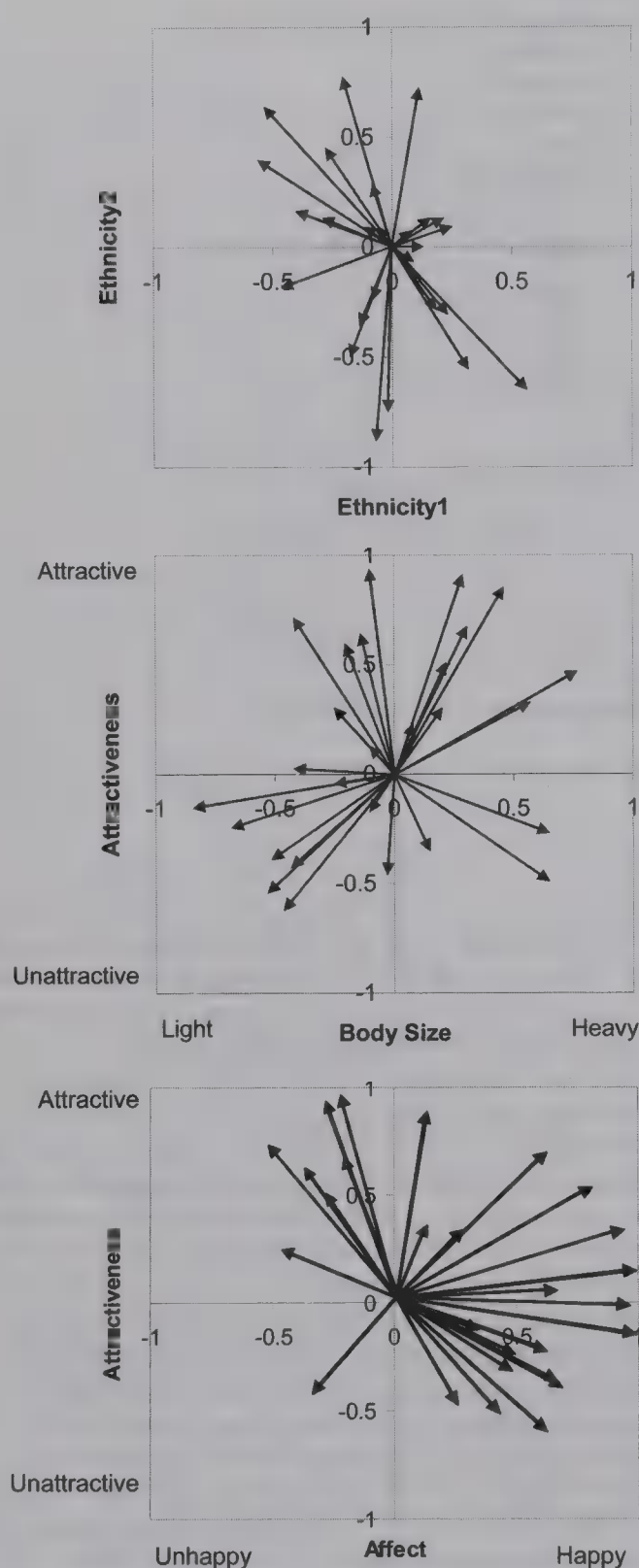


Figure 6. PREFMAP representation of 27 teachers' decision making in the "principal note" context.

dimension could result in a mean signed value of zero. Thus, Table 1 summarizes the preferred directions of the 27 teachers who showed a strong preference for any of the five dimensions in each of the four contexts; a strong preference was indicated by an average utilization value greater than or equal to .50 for a dimension. As demonstrated in the prior analysis of the utilization data, far more teachers utilized affect and attractiveness than the remaining three dimensions. This supplemental analysis reveals that

teachers who utilized these dimensions strongly in their decision making showed a marked preference for happy and attractive girls. Chi-square analyses demonstrated that the preferred direction varied significantly for both the affect and the attractiveness dimensions when collapsed across context,  $\chi^2(1, N = 27) = 23.113$  and 17.894, respectively,  $ps < .05$ . In contrast, teachers who utilized the remaining three dimensions did not show a significant directional preference, although our power to detect directional preferences was quite small for these three attributes, given their limited utilization.

*Individual differences in dimensional utilization and preferred directions.* Nonparametric correlational methods were used to examine individual differences in dimensional utilization and preferred directions. Neither dimensional utilization nor the preferred directions along those dimensions were related to teachers' age, years of teaching experience, teacher satisfaction, and amount of experience with students of diverse backgrounds. In addition, teachers who reported avoiding the use of student characteristics did not utilize these characteristics significantly less than teachers who did not report such avoidance. One-tailed  $t$  tests were used to evaluate whether teachers who reported avoiding body size or ethnicity when making decisions actually used body size or ethnicity significantly less than the remaining teachers. No evaluations were significant. In fact, the average effect size (Cohen's  $d$ ) across the four contexts for utilization of body size by teachers who did and did not report avoiding body size was  $-0.67$ , opposite of the predicted direction of the effect, indicating that teachers who reported avoiding body size actually were more likely to use it. Across the four contexts, utilization of body size correlated with avoidance of using body size .22, .10,  $-.07$ , and .23, with positive correlations indicating that teachers who reported avoiding body size actually used body size information more in decision making. Correlations between avoidance of ethnicity information and use of ethnicity in decision making were similar in magnitude and direction. The average effect sizes comparing the utilization of ethnicity for teachers who did and did not report avoiding the use of ethnicity were .09 and .03 for the first and second ethnicity dimensions, respectively, across the four decision-making contexts. Overall, these findings suggest that teachers were unsuccessful in modifying their decision making in accordance with their desires to avoid use of ethnicity and body size, presumably in part because of the relatively brief stimulus presentation times and the urging to respond quickly.

## Discussion

Previous research suggests that overweight children and some ethnic minority students struggle both socially and academically in school, but few studies have considered the role of teachers' cognitive processing of students' attributes as potential correlates of these difficulties (Puhl & Brownell, 2001, 2004; U.S. Department of Education, National Center for Educational Statistics, 2001). The purpose of this study was to examine two aspects of cognitive processing of person attributes in classroom-relevant contexts—attention and decision making—by investigating teachers' perception of students across multiple stigmatized attributes. We examined teachers' relative attention to competing student-specific characteristics and teachers' utilization of and preference for student characteristics when making different decisions in

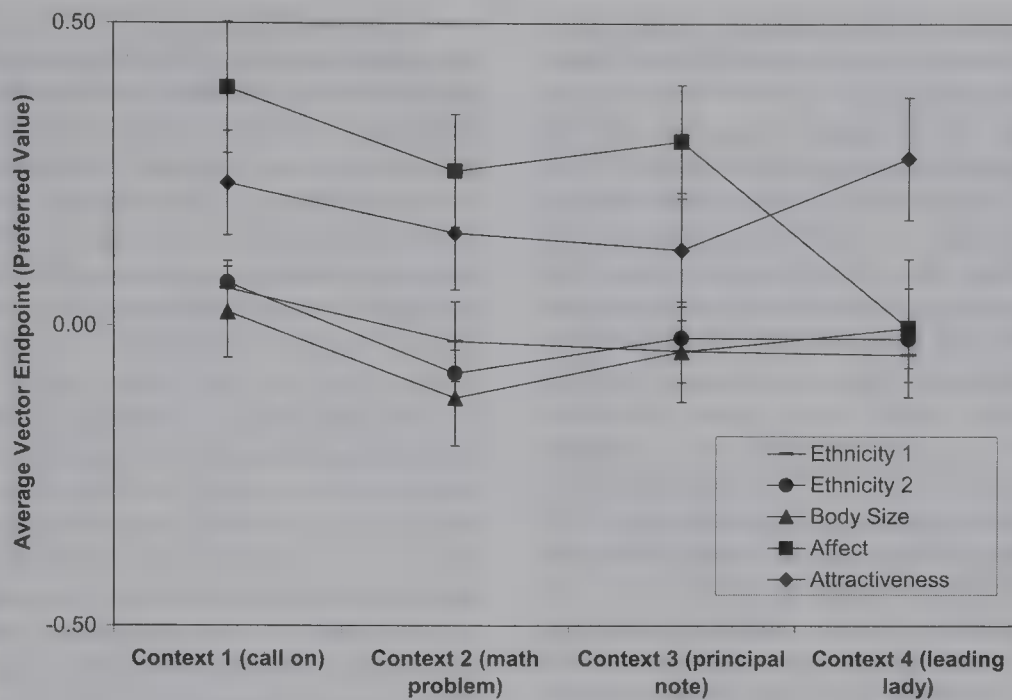


Figure 7. Average preferred directions for five attributes in four contexts. Vertical rules depict standard errors of the means.

classroom-relevant contexts. The student characteristics of particular interest to us in the study were a subset of those found previously to be associated with student performance or teacher judgments—ethnicity, attractiveness, affect, and body size—but had not been investigated relative to one another. Our study emphasized internal validity at the cost of external validity, allowing us to examine the independent impact of these attributes, which often covary under naturally occurring conditions, on teachers' cognitive processing in specific decision-making contexts. Isolating a subset of dimensions of interest also ensured that we had ruled out the influence of other potential covariates of these dimensions (e.g., students' personality, background information) on teachers' processing in these contexts. We recognize that there are many differences between the situations in the study and those that occur in the classroom and that the research program increas-

ingly needs to include stimulus features that make the situations as real as possible. Such standardized situations and stimuli were necessary in this initial study, however, to parse out the extent to which teachers were processing specific student information in classroom-relevant contexts involving rapid decision making. We hypothesized that teachers would show substantial attention to and utilization of students' body size and ethnicity and would exhibit preferences for lighter, nonminority students.

#### Teacher Attention to Student Characteristics

WMDS analysis of the similarity-rating data revealed that teachers attended, as expected, to four interpretable, underlying dimensions that corresponded to the dimensions we built into the stimulus set: body size, ethnicity (two dimensions), and affect. As a

Table 1  
Number of Teachers Showing Strong Preferred Directions for the Five Attributes in the Four Decision-Making Contexts

Attribute	Endpoint	Context				Total
		1	2	3	4	
Ethnicity 1	+ values	1	3	1	0	5
	– values	0	2	1	1	4
Ethnicity 2	+ values	1	2	3	0	6
	– values	0	3	3	2	8
Body size	Heavy	4	3	4	3	14
	Light	1	5	4	1	11
Affect	Happy	16	10	11	7	44
	Sad	3	0	0	6	9
Attractiveness	High	10	8	8	12	38
	Low	2	2	2	3	9

Note. Average dimensional utilization was greater than .50. Context 1 refers to selection of a student who asks a question in class, Context 2 refers to choosing a student to take a note to the principal's office, Context 3 refers to selecting a student to solve a difficult math problem, and Context 4 involves choosing a student to be the leading lady in a classroom skit.



group, participants attended to these dimensions at similar levels, which indicates that the dimensions were perceived to be of similar salience in this particular stimulus set. Attractiveness did not emerge as an independent underlying dimension but correlated moderately with two other dimensions in the WMDS solution. Substantial individual differences in teacher attention to different dimensions emerged, particularly with respect to body size and one of the ethnicity dimensions. Variability in teachers' attentional patterns did not correlate with the individual-differences variables that we assessed, however, such as teacher demographic information and teaching experience with diverse groups of students. Other individual-differences variables, such as teacher ethnicity or body size, could not be examined as moderators of teachers' attention or utilization patterns because of the restricted range of variation but are important to examine in future research.

The use of MDS techniques allowed us to map teachers' perceptual representations of complex person attributes, such as ethnicity. Rather than assuming that ethnicity was perceived as varying continuously along a single dimension, we conducted MDS analyses that indicated that teachers' perceptions of ethnicity were best represented nominally along two dimensions. Girls from different ethnic groups were perceived as belonging to one of three fairly distinct clusters: Asian American, African American, and Caucasian. Dark-haired Caucasian or possibly Hispanic stimuli served as "transition" stimuli between the groups. These results raise questions for future research in terms of how ethnicity is perceived and how its perception might differ as a function of the individuals' own ethnicity and the social context in which the perceiver exists. Implicit, performance-based measures may allow us to capture more accurately how individuals perceive such complex person attributes under specific conditions of interest.

### *Teacher Utilization of Student Characteristics in Decision Making*

Given the documented academic and social difficulties of overweight and ethnic-minority students, we hypothesized that teachers would rely heavily on students' body size and ethnicity when selecting students in four contexts that teachers had deemed classroom relevant: choosing a student to call on in class, selecting a student to take a note to the principal's office, assigning a student to complete a difficult math problem on the board, and picking a student to take a leading-lady role in a class skit. Although the length of time that stimuli were presented may seem brief, teachers routinely make similar rapid decisions in the classroom, and the scaling analyses indicated that teachers systematically used multidimensional student information to make decisions in the preference-ratings task. Unexpectedly, PREFMAP analyses of teachers' choice data indicated that teachers relied predominantly on girls' affect and attractiveness in these decision-making contexts and used ethnicity and body size to a much more limited extent. Despite the moderate correlation between attractiveness and body size, a different pattern of findings emerged for teacher utilization of girls' body size versus attractiveness. Thus, it appears that teachers attended to but did not utilize body-size and ethnicity information when making decisions in classroom contexts. We need to conduct future work to better understand the extent to which the decisions made under these highly controlled conditions converge with decision making in actual classroom contexts.

Teachers' surprisingly infrequent use of body-size and ethnicity information warrants further investigation. Their efforts to present themselves in a favorable light provide a poor account of this finding; the tasks were speeded and required teachers to make over 1,000 choices between pairs of complex multidimensional stimuli, which presumably decreased the likelihood that participants would or could avoid the use of body size or ethnicity when making decisions. Additionally, teachers who reported attempting not to use this information nonetheless used it to a similar degree as other teachers. Future research should examine whether teachers rely on students' body size and ethnicity when making decisions in other situations that were not presented in our study. For example, teachers might make greater use of this information when making decisions that necessitate excluding, rather than including, children (e.g., selecting someone to sit out of an activity). Finally, our findings raise the possibility that other key players in the overweight or ethnic-minority child's social and academic environment are more central to their difficulties. Peers, for example, play an important role in children's social and academic environments at school. Although teachers may not use body size and ethnicity as a basis for student selection, one can readily imagine that some children use these criteria when choosing a teammate in gym class or a study partner for a science project.

Teachers' heavy reliance on attractiveness and affect across decision-making contexts stands in stark contrast to their limited use of body size and ethnicity. Previous researchers have minimized variation in body size when investigating the influence of student attractiveness on teachers' attitudes and behaviors, by using only head shots of students or depicting only average-weight students. In contrast, we allowed both characteristics to vary substantially in our photo stimulus set and minimized the relationship between body size and attractiveness to the extent possible when selecting stimuli. By tightly controlling the covariation between the body size and the attractiveness of the stimuli, we were able to distinguish between the patterns of results for attractiveness compared to body size. The divergence in the findings for attractiveness and body-size information in the present study indicates that aspects of attractiveness that are unrelated to body size may exert a significant influence on teacher decision making. Facial affect was a significant correlate of, though not synonymous with, attractiveness. Some teachers chose happy and attractive children in the decision-making tasks, while others preferred those who were smiling, regardless of attractiveness, or highly attractive but unhappy-looking girls.

The impact of body size and affect on attractiveness raises questions about what exactly attractiveness is and what aspects of attractiveness influence teacher decisions. Perhaps attractiveness consists of multiple components that vary in stability, such that some aspects routinely suggest attractiveness (e.g., body size), whereas other aspects (e.g., emotional expression) may be more malleable and contextually bound. Regardless, teachers' uniformly heavy reliance on attractiveness when selecting students for common classroom tasks is discouraging and suggests the importance of intervention efforts designed to minimize this influence. Future research examining the construct of attractiveness is important to further understand why it is such a reliably important variable to teacher expectations and behaviors, independent from facial affect and body size.



Some variability in teachers' utilization of ethnicity, body size, affect, and attractiveness emerged, but the overwhelming majority of teachers used affect and attractiveness far more than the other dimensions. Given the homogeneity of the parameter estimates, it was unsurprising that teachers' individual differences in utilization patterns were unrelated to the assessed individual-differences variables. In sum, PREFMAP analyses of teachers' choice data demonstrated that teachers relied primarily on students' affect and attractiveness when making decisions in common classroom contexts.

### *Teachers' Preferred Directions of Utilized Student Characteristics in Decision Making*

When examining teachers' preferred directions along heavily utilized dimensions, we anticipated that teachers not only would rely heavily on body size and ethnicity when making decisions but also would exhibit strong preferences for lighter, nonminority students. Few teachers displayed strong utilization of body size and ethnicity, however, and these teachers' preferred directions were distributed similarly across the two ends of the dimensions. In other words, teachers did not display systematic preferences for light students over heavy students or Caucasian or Asian American students over African American students. These findings are inconsistent with the marked social and academic difficulties reported by heavier and ethnic-minority students, and they suggest the importance of examining teacher decision making in other contexts (e.g., exclusionary situations) and investigating peers' decision making in school contexts. Further examining the cognitive processes involved in person perception may yield greater understanding of how stigmatized information is used when varying amounts of information are known about a person. For instance, does an individual's ethnicity become less important to person perception and to decisions based on these perceptions as additional person information becomes available? Such information about how teachers process ethnicity information relative to other important student information could raise questions about real-world interventions, such as the training offered by many school districts designed to enhance teachers' cultural sensitivity to and awareness of ethnically based biases.

Almost all of the teachers who showed strong utilization of affect and attractiveness routinely selected happy-looking, attractive children, although a few teachers systematically chose sad-looking or unattractive children. This overwhelming preference for attractive children is cause for concern and consideration of efforts to diminish this effect, as the stable aspects of attractiveness do not necessarily imply intelligence, responsibility, likability, or academic potential any more than body size or ethnicity do.

Our efforts to enhance the internal validity of our conclusions necessarily constrained the ecological validity of our findings. Thus, one should use caution when generalizing these findings to different student and teacher populations, experimental and stimulus conditions, and decision-making contexts. For instance, we included only girls in the stimulus set (i.e., we did not include gender as a dimension in the stimulus set) to constrain the dimensions being investigated to a tractable number, and the extent to which similar findings would emerge for teacher processing of boys is unknown. Similarly, although we did not find any processing differences across decision-making contexts, such differences

might emerge in other situations or if different students attributes were examined. We also purposely limited the amount of time that teachers had to examine the photo stimuli in an effort to create an experimental situation that was as analogous as possible to ones in which teachers must make rapid decisions, but we recognize that teachers often have longer to deliberate in many other significant classroom situations. Therefore, the present study was quite limited in its external validity, relative to the abundance of rich observational studies in this area, and future research should examine the convergence of the findings in the present study with observed teacher behavior in classroom situations.

### *Implications and Future Directions*

The present study relied on well-developed methods drawn from cognitive science to investigate the importance of several competing, well-researched, stigmatized characteristics to teachers' perceptions and decision making in common classroom-relevant situations. Using a study design that placed a premium on internal validity while incorporating significant externally valid features, we aimed to better understand teachers' processing of students in controlled decision-making contexts in an effort to supplement the existing research in educational psychology that has used other study designs, such as observational methods. Teachers attended to body size, ethnicity, affect, and attractiveness but unexpectedly relied primarily on affect and attractiveness when selecting students in four common classroom contexts. Teachers also exhibited a marked preference for happy, attractive girls across contexts in this study. One implication of these findings is that teachers may be inadvertently communicating that happy or attractive students are more capable or responsible than students who are have less positive affect or are less attractive. The current work suggests that it may be useful for teachers to attend more to how they are selecting students in these situations involving rapid classroom decision making.

This study is intended to be only a first step in a program of research aimed at understanding teachers' attention to and decision-making processes involving students, particularly those from traditionally stigmatized groups. Future work in this line of research can and should integrate other real-world variables of interest to educators. For instance, researchers should examine teacher decision making in a wider array of classroom situations, as well as other academically relevant situations and social situations. In addition, other theoretically important student information, such as the previous year's school performance, could be added in a standardized fashion to the stimulus set via short descriptive phrases at the bottom of each stimulus photo to increase ecological validity in a standardized fashion. Another future direction entails conducting the same study with older children depicted in photographs, because research indicates that the stigma associated with being overweight increases as children age and that older children are viewed as more responsible for their own condition. Moreover, teachers' striking tendency not to select unattractive children warrants further consideration, as this may exacerbate perceptions of unattractive children as less popular, intelligent, and friendly (Langlois et al., 2000). Finally, the relationship between teacher processing patterns and other teacher characteristics should continue to be investigated. For instance, teachers' own concerns about body shape and weight might



heighten their attention to variability in students' body size. Similarly, ethnic-minority teachers might process student ethnicity information differentially.

Better understanding how teachers think about and make decisions in the classroom, beginning with how they use student information to inform their decisions, may have implications for improving student learning and teacher training. For example, effective teachers may be systematically processing classroom information differently than less effective teachers. Although this study focused on how teachers attend to and utilize four student characteristics, similar work could be conducted examining other complex teacher behaviors that are important to student education, such as problem solving and communication, to understand where in their processing more versus less effective teachers differ. Having a better understanding of the way successful teachers process student information and recognizing the effective components in successful teachers' information processing could be instructive to other teachers by leading to increasingly specific, effective training for new teachers.

Experimental research also could help elucidate how individual differences among teachers impact teaching behaviors and student learning. In the present study, participants reported about some aspects of their teaching experience and various philosophies about teaching, such as whether all children have the potential to succeed and whether teachers should give more attention to struggling children. Although there were no systematic relationships of significance between these variables and which student characteristics teachers attended to or used in the classroom-relevant contexts in this study, there seems to be great potential for examining more teacher variables across different aspects of cognitive processing and teaching behaviors in future work. Experimental work could help determine, for instance, how teachers who believe that all children have the potential to succeed differ in their teaching style compared to those who do not. Such egalitarian teachers may be more likely to attend to negative facial affect, and they also may engage in a whole host of other thoughts or behaviors that are distinct from teachers who believe otherwise.

More generally, the present work highlights the feasibility and potential utility of translating performance-based paradigms drawn from cognitive science to investigate individual differences in complex social perception, even though these methods traditionally have been used to study normative processing of simple, highly artificial stimuli (Treat et al., 2001, 2002, 2007; Viken et al., 2002). It also serves to illustrate how highly controlled experimental work can be used to complement more externally valid research to better understand the processes underlying the behavioral phenomenon of interest. In particular, the similarity- and preference-ratings paradigms, in conjunction with scaling techniques, should be used more widely to investigate the role of attention and decision-making processes in more complex social phenomena, such as stigma and prejudice, which are particularly important to investigate in educational contexts. These paradigms allow for the simultaneous examination of multiple person attributes, which is more consistent with what perceivers must do in the real world. They also allow for greater experimental control and clearer interpretation of results than studies that maximize external validity, such as those using observational designs. Most important, investigating specific stages in information processing using such performance-based paradigms will provide a more

detailed understanding of how person features are processed and used in decision making, which may enhance the effectiveness of teacher training workshops.

## References

- Ambady, N., Hallahan, M., & Conner, B. (1999). Accuracy of judgments of sexual orientation from thin slices of behavior. *Journal of Personality and Social Psychology*, 77, 538–547.
- Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, 111, 256–274.
- Ambady, N., & Rosenthal, R. (1993). Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. *Journal of Personality and Social Psychology*, 64, 431–441.
- Barab, S. A., Hay, K. E., & Yamagata-Lynch, L. C. (2001). Constructing networks of activity: An in-situ research methodology. *Journal of the Learning Sciences*, 10, 63–112.
- Barab, S. A., & Plucker, J. A. (2002). Smart people or smart contexts? Cognition, ability, and talent development in an age of situated approaches to knowing and learning. *Educational Psychologist*, 37, 165–182.
- Bechtel, G. G., Tucker, L. R., & Chang, W. (1971). A scalar product model for the multidimensional scaling of choice. *Psychometrika*, 36, 369–388.
- Benson, P. L., Severs, D., Tatgenhorst, J., & Loddengaard, N. (1980). The social costs of obesity: A non-reactive field study. *Social Behavior and Personality*, 8, 91–96.
- Blatchford, P. (2003). A systematic observational study of teachers' and pupils' behaviour in large and small classes. *Learning and Instruction*, 13, 569–595.
- Borg, I., & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications*. New York: Springer-Verlag.
- Borkenau, P., Mauer, N., Riemann, R., Spinath, F. M., & Angleitner, A. (2004). Thin slices of behavior as cues of personality and intelligence. *Journal of Personality and Social Psychology*, 86, 599–614.
- Canning, H., & Mayer, J. (1966). Obesity—Its possible effect on college acceptance. *New England Journal of Medicine*, 275, 1172–1174.
- Carroll, J. D. (1972). Individual differences and multidimensional scaling. In R. N. Shepard, A. K. Romney, & S. B. Nerlove (Eds.), *Multidimensional scaling: Vol. 1: Theory* (pp. 105–155). New York: Seminar Press.
- Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35, 238–319.
- Carroll, J. D., & Wish, M. (1974). Models and methods for three-way multidimensional scaling. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology: Vol. 2. Measurement, psychophysics, and neural information processing* (pp. 57–105). New York: Freeman.
- Chang, D. F., & Sue, S. (2003). The effects of race and problem type on teachers' assessments of student behavior. *Journal of Consulting and Clinical Psychology*, 71, 235–242.
- Chang, J. J., & Carroll, J. D. (1972). *How to use PREFMAP and PREFMAP-2—Programs which relate preference data to multidimensional scaling solutions*. Unpublished manuscript, Bell Telephone Labs, Murray Hill, NJ.
- Cloutier, J., Mason, M. F., & Macrae, C. N. (2005). The perceptual determinants of person construal: Reopening the social-cognitive toolbox. *Journal of Personality and Social Psychology*, 88, 885–894.
- Davison, M. L. (1992). *Multidimensional scaling*. Malabar, FL: Krieger Education Trust.
- Education Trust. (2004). *Education watch: The nation: Key education facts and figures: Achievement, attainment, and opportunity from elementary school through college*. Retrieved June 28, 2007, from <http://www2.edtrust.org/edtrust/summaries2004/USA.pdf>
- Fazio, R. H., & Dunton, B. C. (1997). Categorization by race: The impact

- of automatic and controlled components of racial prejudice. *Journal of Experimental Social Psychology*, 33, 451–470.
- Fiske, S. T. (1993). Social cognition and social perception. *Annual Review of Psychology*, 44, 155–194.
- Friedman, H. S., DiMatteo, M. R., & Mertz, T. I. (1980). Nonverbal communication on television news: The facial expression of broadcasters during coverage of a presidential election campaign. *Personality and Social Psychology Bulletin*, 6, 427–435.
- Goldin-Meadow, S. (2004). Gesture's role in the learning process. *Theory Into Practice*, 43, 314–321.
- Gortmaker, S. L., Must, A., Perrin, J. M., Sobol, A. M., & Dietz, W. H. (1993). Social and economic consequences of overweight in adolescence and young adulthood. *New England Journal of Medicine*, 329, 1008–1012.
- Graham, S. (1990). Communicating low ability in the classroom: Bad things good teachers sometimes do. In S. Graham & V. S. Folkes (Eds.), *Attribution theory: Applications to achievement, mental health, and interpersonal conflict*. Applied social psychology (pp. 17–36). Northvale, NJ: Erlbaum.
- Hebl, M. R., & Mannix, L. M. (2003). The weight of obesity in evaluating others: A mere proximity effect. *Personality and Social Psychology Bulletin*, 29, 28–38.
- Jussim, L. (1989). Teacher expectations: Self-fulfilling prophecies, perceptual biases, and accuracy. *Journal of Personality and Social Psychology*, 57, 469–480.
- Jussim, L. (1991). Social perception and social reality: A reflection-construction model. *Psychological Review*, 98, 54–73.
- Kumar, S. (2002). WinMorph [computer software]. Retrieved September 25, 2001, from <http://www.debugmode.com/winmorph/download.php>
- Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin*, 126, 390–423.
- Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology*, 51, 93–120.
- Macrae, C. N., & Bodenhausen, G. V. (2001). Social cognition: Categorical person perception. *British Journal of Psychology*, 92, 239–255.
- Macrae, C. N., Quinn, K. A., Mason, M. F., & Quadflieg, S. (2005). Understanding others: The face and person construal. *Journal of Personality and Social Psychology*, 89, 686–695.
- McFall, R. M. (1982). A review and reformulation of the concept of social skills. *Behavioral Assessment*, 4, 1–33.
- McFall, R. M. (1990). The enhancement of social skills: An information-processing analysis. In W. L. Marshall, & D. R. Laws (Eds.), *Handbook of sexual assault: Issues, theories, and treatment of the offender*. Applied clinical psychology (pp. 311–330). New York: Plenum Press.
- Meulman, J., Heiser, W. J., & Carroll, J. D. (1986). *PREFMAP-3 user's guide*. Unpublished manuscript, Bell Telephone Labs, Murray Hill, NJ.
- Michel, C., Caldara, R., & Rossion, B. (2006). Same-race faces are perceived more holistically than other-race faces. *Visual Cognition*, 14, 55–73.
- Mottet, T. P., & Richmond, V. P. (2000, November). *Student nonverbal communication and its influence on teachers and teaching: A review of literature*. Paper presented at the meeting of the National Communication Association, Seattle, WA.
- Neal, L. I., McCray, A. D., Webb-Johnson, G., & Bridgest, S. T. (2003). The effects of African American movement styles on teachers' perceptions and reactions. *Journal of Special Education*, 37, 49–57.
- Neumark-Sztainer, D., Falkner, N., Story, M., Perry, C., Hannan, P. J., & Mulert, S. (2002). Weight-teasing among adolescents: Correlations with weight status and disordered eating behaviors. *International Journal of Obesity*, 26, 123–131.
- Nosofsky, R. M. (1992). Similarity scaling and cognitive process models. *Annual Review of Psychology*, 43, 25–53.
- O'Connor, E. A., Fish, M. C., & Yasik, A. E. (2004). The influence of teacher experience on the elementary classroom system: An observational study. *Journal of Classroom Interaction*, 39, 11–18.
- Puhl, R., & Brownell, K. D. (2001). Bias, discrimination, and obesity. *Obesity Research*, 9, 788–805.
- Puhl, R., & Brownell, K. D. (2004). Bias, discrimination, and obesity. In G. A. Bray & C. Bouchard (Eds.), *Handbook of obesity: Clinical applications* (2nd ed., pp. 69–74). New York: Marcel Dekker.
- Ritts, V., Patterson, M. L., & Tubbs, M. E. (1992). Expectations, impressions, and judgments of physically attractive students: A review. *Review of Educational Research*, 62, 413–426.
- Rosenthal, R., & Jacobson, L. (1966). Teachers' expectancies: Determinants of pupils' IQ gains. *Psychological Report*, 19, 115–118.
- Roth, W. (1998). Situated cognition and assessment of competence in science. *Evaluation and Program Planning*, 21, 155–169.
- Schiffman, S. S., Reynolds, M. L., & Young, F. W. (1981). *Introduction to multidimensional scaling: Theory, methods, and applications*. Orlando, FL: Academic Press.
- Strauss, R. S. (2002). Childhood obesity. *Pediatric Clinics of North America*, 49, 175–201.
- Stuhlman, M. W., & Pianta, R. C. (2002). Teachers' narratives about their relationships with children: Associations with behavior in classrooms. *School Psychology Review*, 31, 148–163.
- Tauber, R. T. (1998). *Good or bad, what teachers expect from students they generally get* (Report EDO-SP-97-7). Washington, DC: ERIC Clearinghouse.
- Treat, T. A., McFall, R. M., Viken, R. J., & Kruschke, J. K. (2001). Using cognitive science methods to assess the role of social information processing in sexually coercive behavior. *Psychological Assessment*, 13, 549–565.
- Treat, T. A., McFall, R. M., Viken, R. J., Kruschke, J. K., Nosofsky, R. M., & Wang, S. S. (2007). Clinical cognitive science: Applying quantitative models of cognitive processing to examination of cognitive aspects of psychopathology. In R. W. J. Neufeld (Ed.), *Advances in clinical cognitive science: Formal modeling and assessment of processes and symptoms* (pp. 179–205). Washington, DC: American Psychological Association.
- Treat, T. A., McFall, R. M., Viken, R. J., Nosofsky, R. M., MacKay, D. B., & Kruschke, J. K. (2002). Assessing clinically relevant perceptual organization with multidimensional scaling techniques. *Psychological Assessment*, 14, 239–252.
- U.S. Department of Education, National Center for Educational Statistics. (2001). *Educational achievement and Black-White inequality* (NCES 2001-061). Washington, DC: U.S. Government Printing Office.
- Viken, R. J., Treat, T. A., Nosofsky, R. M., McFall, R. M., & Palmeri, T. (2002). Modeling individual differences in perceptual and attentional processes related to bulimic symptoms. *Journal of Abnormal Psychology*, 111, 598–609.
- Zebrowitz, L. A., Hall, J. A., Murphy, N. A., & Rhodes, G. (2002). Looking smart and looking good: Facial cues to intelligence and their origins. *Personality and Social Psychology Bulletin*, 28, 238–249.

Received November 3, 2005

Revision received June 28, 2007

Accepted July 3, 2007 ■



The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

**Manuscript preparation.** Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (5th ed.). Manuscripts may be copyedited for bias-free language (see chap. 2 of the *Publication Manual*). Formatting instructions (all copy must be double-spaced) and instructions on the preparation of tables, figures, references, metrics, and abstracts appear in the *Manual*. For APA's Checklist for Manuscript Submission, see [www.apa.org/journals](http://www.apa.org/journals). **Abstract and keywords.** All manuscripts must include an abstract containing a maximum of 180 words typed on a separate page. After the abstract, please supply up to five keywords or brief phrases. **References.** References should be listed in alphabetical order. Each listed reference should be cited in text, and each text citation should be listed in the References. Basic formats are as follows:

- Haag, L., & Stern, E. (2003). In search of the benefits of learning Latin. *Journal of Educational Psychology*, 95, 174–178.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Johnson, D. W., & Johnson, R. T. (1990). Cooperative learning and achievement. In S. Sharon (Ed.), *Cooperative learning: Theory and research* (pp. 173–202). New York: Praeger.

Adequate description of participants is critical to the science and practice of educational psychology; this allows readers to assess the results, determine generalizability of findings, and make comparisons in replications, extensions, literature reviews, or secondary data analyses. Authors should see guidelines for sample-subject description in the *Manual*. Appropriate indexes of effect size or strength of relationship should be incorporated in the results section of the manuscript (see pp. 5, 25–26 of the *Manual*). Information that allows the reader to assess not only the significance but also the magnitude of the observed effects or relationships clarifies the importance of the findings. **Figures.** Graphics files are welcome if supplied as Tiff, EPS, or PowerPoint. High-quality printouts or glossies are needed for *all* figures. The minimum line weight for line art is 0.5 point for optimal printing. When possible, please place symbol legends below the figure image instead of to the side. Original color figures can be printed in color at the editor's and publisher's discretion provided the author agrees to pay \$255 for one figure, \$425 for two figures, \$575 for three figures, \$675 for four figures, and \$55 for each additional figure.

**Publication policies.** APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. APA policy regarding posting articles on the Internet may be found at [www.apa.org/journals](http://www.apa.org/journals). In addition, it is a violation of APA Ethical Principles to publish "as original data, data that have been previously published" (Standard 8.13). As this is a primary journal that publishes original material only, APA policy prohibits publication of any manuscript or data that have already been published in whole or substantial part elsewhere. Authors have an obligation to consult journal editors concerning prior publication of any data on which their article depends. In addition, APA Ethical Principles specify that "after research

results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release" (Standard 8.14). Authors must have available their data throughout the editorial review process and for at least 5 years after the date of publication.

**Masked review policy.** Because the *Journal* has a masked review policy, authors submitting manuscripts are required to include, with each copy of the manuscript, a cover sheet that shows the title of the manuscript, the authors' names and institutional affiliations, the date the manuscript is submitted, and footnotes identifying the authors or their affiliations. The first page of the manuscript should omit the authors' names and affiliations but should include the title of the manuscript and the date it is submitted. Every effort should be made by the authors to see that the manuscript itself contains no clues to their identities. Authors are required to state in writing that they have complied with APA ethical standards in the treatment of their sample, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained at [www.apa.org/ethics/](http://www.apa.org/ethics/) or by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242. APA requires authors to reveal any possible conflict of interest in the conduct and reporting of research (e.g., financial interests in a test procedure, funding by pharmaceutical companies for drug research). Authors of accepted manuscripts will be required to transfer copyright to APA.

**Permissions.** Authors of accepted papers are required to obtain and provide to the editor on final acceptance all necessary permissions to reproduce in print and electronic form any copyrighted work, including, for example, test materials or portions thereof and photographs of people.

**Supplemental materials.** APA can now place supplementary materials online, which will be available via the published article in the PsycARTICLES database. To submit such materials, please see [www.apa.org/journals/authors/suppmaterial.html](http://www.apa.org/journals/authors/suppmaterial.html) for details. Authors of accepted papers will be asked to work with the editor and production staff to provide supplementary materials as appropriate.

**Submission.** Authors should submit their manuscripts electronically via the Manuscript Submission Portal at [www.apa.org/journals/edu](http://www.apa.org/journals/edu) (follow the link "Submit Manuscripts Electronically"). A checklist for manuscript submission, including guidelines for preparing the electronic file, can be found at [www.apa.org/journals/](http://www.apa.org/journals/). Correspondence regarding manuscripts should be sent to the Editor, Art Graessen, University of Memphis, Journal of Educational Psychology, 202 Psychology Building, Memphis, TN 38152-3230. In addition to addresses and phone numbers, authors should supply e-mail addresses, as most communications will be by e-mail. Fax numbers, if available, should also be provided for potential use by the editorial office and later by the production office. Authors should keep a copy of the manuscript to guard against loss. E-mail correspondence may be addressed to [jedgar@memphis.edu](mailto:jedgar@memphis.edu).

**Preparing files for production.** If your manuscript is accepted for publication, please follow the guidelines for file formats and naming provided at [www.apa.org/journals/authors/preparing\\_files.html](http://www.apa.org/journals/authors/preparing_files.html). If your manuscript was mask reviewed, please ensure that the final version for production includes a byline and full author note for typesetting.





# Charles C Thomas

PUBLISHER • LTD.

P.O. Box 19265  
Springfield, IL 62794-9265

## Book Savings

(on separate titles only)

Save 10% on 1 Book!  
Save 15% on 2 Books!  
Save 20% on 3 Books!

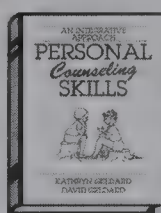
- Brock, Michael G. and Samuel Saks—**CONTEMPORARY ISSUES IN FAMILY LAW AND MENTAL HEALTH.** '08, 160 pp. (7 x 10), paper.

- Brooke, Stephanie L.—**THE USE OF THE CREATIVE THERAPIES WITH SURVIVORS OF DOMESTIC VIOLENCE.** '08, 344 pp. (7 x 10), 57 il., (14 in color).

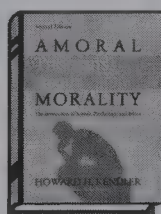
- Coulacoglou, Carina—**EXPLORING THE CHILD'S PERSONALITY: Developmental, Clinical and Cross-Cultural Applications of the Fairy Tale Test.** '08, 330 pp. (8 x 10), 22 il., 41 tables.



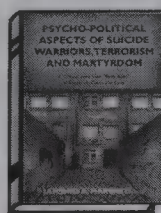
- Brooke, Stephanie L.—**THE CREATIVE THERAPIES AND EATING DISORDERS.** '08, 304 pp. (7 x 10), 20 il., 2 tables, \$64.95, hard, \$44.95, paper.



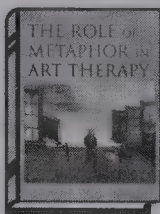
- Geldard, Kathryn & David Geldard—**PERSONAL COUNSELING SKILLS: An Integrative Approach.** '08, 316 pp. (7 x 10), 20 il., 3 tables, \$49.95, paper.



- Kendler, Howard H.—**AMORAL THOUGHTS ABOUT MORALITY: The Intersection of Science, Psychology, and Ethics. (2nd Ed.)** '08, 270 pp. (7 x 10), \$59.95, hard, \$39.95, paper.



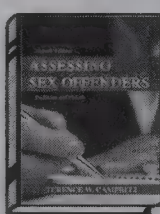
- Marvasti, Jamshid A.—**PSYCHO-POLITICAL ASPECTS OF SUICIDE WARRIORS, TERRORISM AND MARTYRDOM: A Critical View from "Both Sides" in Regard to Cause and Cure.** '08, 374 pp. (7 x 10), \$73.95, hard, \$53.95, paper.



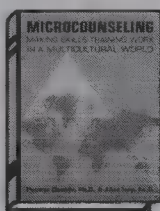
- Moon, Bruce L.—**INTRODUCTION TO ART THERAPY: Faith in the Product. (2nd Ed.)** '08, 226 pp. (7 x 10), 20 il., \$53.95, hard, \$33.95, paper.



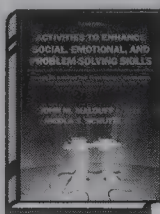
- Arrington, Doris Banowsky—**ART, ANGST, AND TRAUMA: Right Brain Interventions with Developmental Issues.** '07, 278 pp. (7 x 10), 123 il., \$63.95, hard, \$43.95, paper.



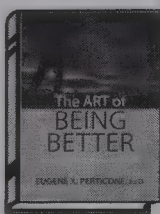
- Campbell, Terence W.—**ASSESSING SEX OFFENDERS: Problems and Pitfalls. (2nd Ed.)** '07, 376 pp. (7 x 10), 46 tables, \$74.95, hard, \$54.95, paper.



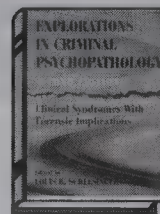
- Daniels, Thomas & Allen Ivey—**MICROCOUNSELING: Making Skills Training Work in a Multicultural World.** '07, 296 pp. (7 x 10), 12 il., 3 tables, \$65.95, hard, \$45.95, paper.



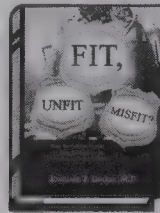
- Malouff, John M. & Nicola S. Schutte—**ACTIVITIES TO ENHANCE SOCIAL, EMOTIONAL, AND PROBLEM-SOLVING SKILLS: Seventy-six Activities That Teach Children, Adolescents, and Adults Skills Crucial to Success in Life. (2nd Ed.)** '07, 248 pp. (8 1/2 x 11), 3 il., \$44.95 spiral (paper).



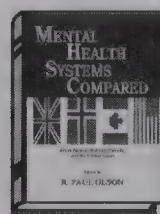
- Perticone, Eugene X.—**THE ART OF BEING BETTER: An Approach to Personal Growth.** '07, 268 pp. (7 x 10), \$58.95, hard, \$38.95, paper.



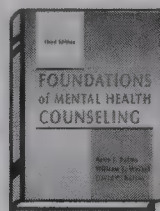
- Schlesinger, Louis B.—**EXPLORATIONS IN CRIMINAL PSYCHOPATHOLOGY: Clinical Syndromes With Forensic Implications. (2nd Ed.)** '07, 432 pp. (7 x 10), 3 il., 10 tables \$79.95, hard, \$55.95, paper.



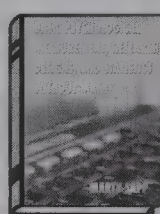
- Decker, Kathleen P.—**FIT, UNFIT OR MISFIT? How To Perform Fitness for Duty Evaluations in Law Enforcement Professionals.** '06, 284 pp. (7 x 10), 3 il., 43 tables, \$61.95, hard, \$41.95, paper.



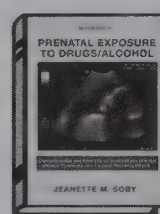
- Olson, R. Paul—**MENTAL HEALTH SYSTEMS COMPARED: Great Britain, Norway, Canada, and the United States.** '06, 398 pp. (8 1/2 x 11), 10 il., 65 tables, \$89.95, hard, \$64.95, paper.



- Palmo, Artis J., William J. Weikel & David P. Borsos—**FOUNDATIONS OF MENTAL HEALTH COUNSELING. (3rd Ed.)** '06, 468 pp. (7 x 10), 5 il., 3 tables, \$85.95, hard, \$61.95, paper.



- Sapp, Marty—**BASIC PSYCHOLOGICAL MEASUREMENT, RESEARCH DESIGNS, AND STATISTICS WITHOUT MATH.** '06, 288 pp. (7 x 10), 2 il., 34 tables, \$62.95, hard, \$42.95, paper.



- Soby, Jeanette M.—**PRENATAL EXPOSURE TO DRUGS/ALCOHOL: Characteristics and Educational Implications of Fetal Alcohol Syndrome and Cocaine/Polydrug Effects. (2nd Ed.)** '06, 188 pp. (7 x 10), 7 il., 21 tables, \$44.95, hard, \$28.95, paper.

5 easy ways to order!



PHONE:  
1-800-258-8980  
or (217) 789-8980



FAX:  
(217) 789-9130



EMAIL:  
books@ccthomas.com

Web: www.ccthomas.com



MAIL:  
Charles C Thomas •  
Publisher, Ltd.  
P.O. Box 19265  
Springfield, IL 62794-9265

Complete catalog available at ccthomas.com • books@ccthomas.com

Books sent on approval • Shipping charges: \$7.75 min. U.S. / Outside U.S., actual shipping fees will be charged • Prices subject to change without notice

\*Savings include all titles shown here and on our web site. For a limited time only.

When ordering, please refer to promotional code JEDP0508 to receive your discount.



# NEW RELEASES

from the American Psychological Association



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION



## The I/O Consultant

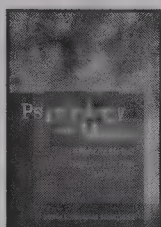
Advice and Insights for Building  
a Successful Career

*Edited by Jerry W. Hedge and Walter C. Borman*

2008. 328 pages. Hardcover.

ISBN 978-1-4338-0339-0 ■ Item # 4316106

List: \$79.95 ■ APA Member/Affiliate: \$49.95



## Psychology As a Major

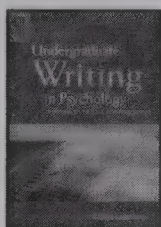
Is It Right for Me and What Can I Do  
With My Degree?

*Donna E. Palladino Schultheiss*

2008. 288 pages. Paperback.

ISBN 978-1-4338-0336-9 ■ Item # 4313016

List: \$29.95 ■ APA Member/Affiliate: \$24.95



## Undergraduate Writing in Psychology

Learning to Tell the Scientific Story

*R. Eric Landrum*

2008. 208 pages. Paperback.

ISBN 978-1-4338-0332-1 ■ Item # 4313015

List: \$29.95 ■ APA Member/Affiliate: \$24.95



## Surviving Graduate School in Psychology

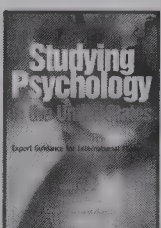
A Pocket Mentor

*Tara L. Kuther*

2008. 344 pages. Paperback.

ISBN 978-1-4338-0346-8 ■ Item # 4313019

List: \$34.95 ■ APA Member/Affiliate: \$29.95



## Studying Psychology in the United States

Expert Guidance for International Students

*Edited by Nadia T. Hasan, Nadya A. Fouad,  
and Carol Williams-Nickelson*

2008. 208 pages. Paperback.

ISBN 978-1-4338-0341-3 ■ Item # 4313017

List: \$29.95 ■ APA Member/Affiliate: \$29.95



## The Collaborative Psychotherapist

Creating Reciprocal Relationships  
With Medical Professionals

*Nancy Breen Ruddy, Dorothy A. Borresen,  
and William B. Gunn, Jr.*

2008. 232 pages. Hardcover.

ISBN 978-1-4338-0338-3 ■ Item # 4317152

List: \$49.95 ■ APA Member/Affiliate: \$39.95



## Cultural Competence in Trauma Therapy

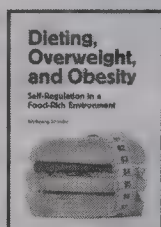
Beyond the Flashback

*Laura S. Brown*

2008. 280 pages. Hardcover.

ISBN 978-1-4338-0337-6 ■ Item # 4317149

List: \$59.95 ■ APA Member/Affiliate: \$49.95



## Dieting, Overweight, and Obesity

Self-Regulation in a Food-Rich Environment

*Wolfgang Stroebe*

2008. 256 pages. Hardcover.

ISBN 978-1-4338-0335-2 ■ Item # 4317148

List: \$59.95 ■ APA Member/Affiliate: \$49.95



## Transcending Self-Interest

Psychological Explorations of the Quiet Ego

*Edited by Heidi A. Wayment and Jack J. Bauer*

2008. 272 pages. Hardcover.

ISBN 978-1-4338-0340-6 ■ Item # 4317153

List: \$79.95 ■ APA Member/Affiliate: \$49.95



## Women Street Hustlers

Who They Are and How They Survive

*Barbara A. Rockell*

2008. 232 pages. Hardcover.

ISBN 978-1-4338-0333-8 ■ Item # 4316104

List: \$69.95 ■ APA Member/Affiliate: \$49.95



## Law and Mental Health Professionals

Massachusetts

Third Edition

*Justice Jonathan Brant*

2008. 288 pages. Hardcover.

ISBN 978-1-4338-0334-5 ■ Item # 4315011

List: \$99.95 ■ APA Member/Affiliate: \$74.95

AD0587

To Order: 800-374-2721 • [www.apa.org/books](http://www.apa.org/books)



# Intervening in Children's Lives

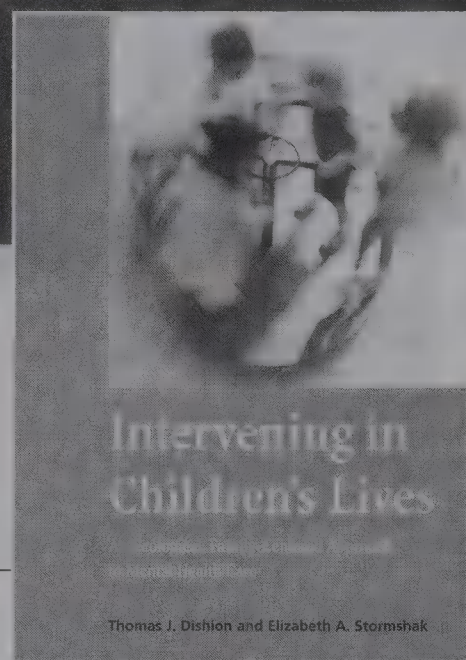
## An Ecological, Family-Centered Approach to Mental Health Care

Thomas J. Dishion and Elizabeth A. Stormshak

**M**ental health interventions for children and adolescents often flow from adult clinical models, which emphasize individual change. Yet, to accomplish long-lasting change for children and adolescents, services need to consider developmental norms, the developmental status of the child or adolescent, and the fact that mental health issues for this population are embedded in family, peer, and sibling relationships. In *Intervening in Children's Lives: An Ecological, Family-Centered Approach to Mental Health Care*, the authors describe a family-centered approach that engages children, adolescents, and their families, leveraging their motivation to change. Never before has there been a comprehensive, systematic framework for linking empirically supported interventions for this clinical population. Useful as both a preventive checkup and a more intensive intervention, this approach may be delivered in schools and other community settings to have the greatest public health impact.

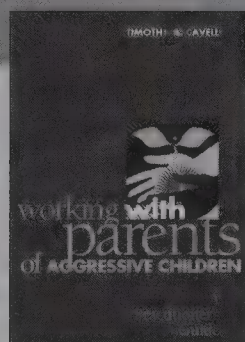
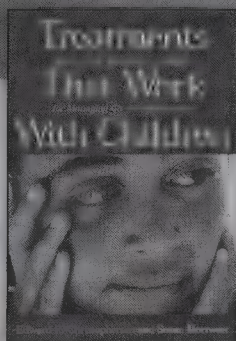
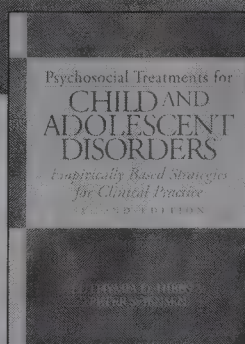
The literature reveals promising findings, in that highest-risk youth are more likely to respond well to ecologically based interventions, and this approach is consistent with others showing long-lasting effects. 2007. 320 pages. Hardcover.

List: \$69.95 ■ APA Member/Affiliate: \$49.95 ■ ISBN 1-59147-428-0 ■ Item # 4317115 ■ ISBN-13: 978-1-59147-428-9



### CONTENTS:

Chapter 1. Child and Family Intervention from an Ecological Perspective: Introduction ■ **Part I. Conceptual Overview** ■ Chapter 2. The Ecology of Development and Change ■ Chapter 3. Family and Peer Social Interaction ■ Chapter 4. The Ecological Family Intervention and Therapy Model ■ **Part II. The Family Check-Up** ■ Chapter 5. Initial Contacts That Establish a Collaborative Set ■ Chapter 6. Ecological Assessment ■ Chapter 7. Mobilizing Change With the Family Check-Up ■ **Part III. Intervention Strategies** ■ Chapter 8. Brief Parenting Interventions ■ Chapter 9. Interventions With Children and Adolescents ■ Chapter 10. Family Management Therapy ■ Chapter 11. Parent Intervention Groups ■ Chapter 12. Child and Adolescent Intervention Groups ■ **Part IV. Professional and Ethical Considerations** ■ Chapter 13. The Ecology of the Child and Family Therapist ■ Chapter 14. Ethical and Professional Standards in Child and Family Interventions



## ALSO AVAILABLE

### WORKING WITH PARENTS OF AGGRESSIVE CHILDREN A Practitioner's Guide

Timothy A. Cavell

2000 ■ 267 pages ■ Hardcover ■ List: \$39.95

APA Member/Affiliate: \$34.95 ■ ISBN 1-55798-637-1

Item # 431631A ■ ISBN-13: 978-1-55798-637-5

### PSYCHOSOCIAL TREATMENTS FOR CHILD AND ADOLESCENT DISORDERS Empirically Based Strategies for Clinical Practice Second Edition

Edited by Euthymia D. Hibbs and Peter S. Jensen

2005 ■ 839 pages ■ Hardcover ■ List: \$69.95 ■ APA Member/Affiliate: \$54.95

ISBN 1-59147-092-7 ■ Item # 4318006 ■ ISBN-13: 978-1-59147-092-2

### TREATMENTS THAT WORK WITH CHILDREN Empirically Supported Strategies for Managing Childhood Problems

Edward R. Christophersen and Susan L. Mortweet

2001 ■ 309 pages ■ Hardcover ■ List: \$39.95 ■ APA Member/Affiliate: \$34.95

ISBN 1-55798-759-9 ■ Item # 431769A ■ ISBN-13: 978-1-55798-759-4

## APA Books Ordering Information

**800-374-2721**

**www.apa.org/books**

In Washington, DC, call: 202-336-5510

TDD/TTY: 202-336-6123 ■ Fax: 202-336-5502

In Europe, Africa, or the Middle East,  
call: 44-207-240-0856



AMERICAN PSYCHOLOGICAL ASSOCIATION



# BEST SELLERS

from the American Psychological Association



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

## The Wisdom of Coaching

Essential Papers in Consulting Psychology  
for a World of Change

*Edited by Richard R. Kilburg  
and Richard C. Diedrich*

2007. 560 pages. Hardcover.

List: \$99.95

APA Member/Affiliate: \$59.95

ISBN 978-1-59147-787-7

Item # 4317125

## A History of Psychology in Autobiography

Volume IX

*Edited by Gardner Lindzey and William M. Runyan*

2007. 408 pages. Hardcover.

List: \$89.95

APA Member/Affiliate: \$59.95

ISBN 978-1-59147-796-9

Item # 4316091

## Getting In

A Step-by-Step Plan for Gaining Admission  
to Graduate School in Psychology

Second Edition

2007. 222 pages. Paperback.

List: \$19.95

APA Member/Affiliate: \$19.95

ISBN 978-1-59147-799-0

Item # 4313012

## Meaning of Others

Narrative Studies of Relationships

*Edited by Ruthellen Josselson, Amia Lieblich,  
and Dan P. McAdams*

2007. 296 pages. Hardcover.

List: \$79.95

APA Member/Affiliate: \$59.95

ISBN 978-1-59147-816-4

Item # 4316092

## The Muscular Ideal

Psychological, Social, and Medical Perspectives

*J. Kevin Thompson and Guy Cafri*

2007. 288 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-1-59147-792-1

Item # 4317127

## Toward a Science of Distributed Learning

*Edited by Stephen M. Fiore and Eduardo Salas*

2007. 288 pages. Hardcover.

List: \$79.95

APA Member/Affiliate: \$59.95

ISBN 978-1-59147-800-3

Item # 4318039

## Assessing Hispanic Clients Using the MMPI-2 and MMPI-A

*James N. Butcher, Jose Cabiya, Emilia Lucio,  
and Maria Garido*

2007. 280 pages. Hardcover.

List: \$69.95

APA Member/Affiliate: \$49.95

ISBN 978-1-59147-924-6

Item # 4317128

## Case Studies in Emotion-Focused Treatment of Depression

A Comparison of Good and Poor Outcome

*Jeanne C. Watson, Rhonda N. Goldman,  
and Leslie S. Greenberg*

2007. 232 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-1-59147-929-1

Item # 4317129

## Inhibition in Cognition

*Edited by David S. Gorfein and Colin M. MacLeod*

2007. 384 pages. Hardcover.

List: \$79.95

APA Member/Affiliate: \$49.95

ISBN 978-1-59147-930-7

Item # 4318042

## Psychology's Interpretive Turn

The Search for Truth and Agency in Theoretical  
and Philosophical Psychology

*Barbara S. Held*

2007. 432 pages. Hardcover.

List: \$79.95

APA Member/Affiliate: \$49.95

ISBN 978-1-59147-925-3

Item # 4316093

## Everyday Creativity and New Views of Human Nature

Psychological, Social,  
and Spiritual Perspectives

*Edited by Ruth Richards*

2007. 328 pages. Hardcover.

List: \$69.95

APA Member/Affiliate: \$49.95

ISBN 978-0-9792125-7-4

Item # 4317134

## An APA LifeTools Book

### How to Write for a General Audience

A Guide for Academics Who Want to  
Share Their Knowledge With the World  
and Have Fun Doing It

*Kathleen A. Kendall-Tackett, PhD*

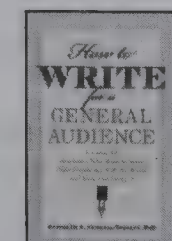
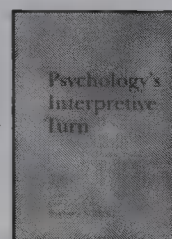
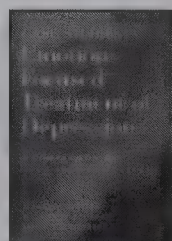
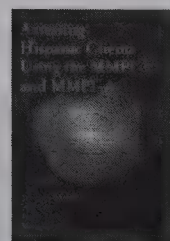
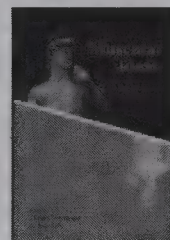
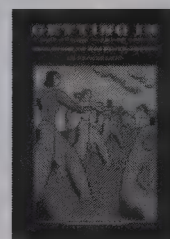
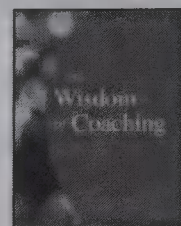
2007. 352 pages. Paperback.

List: \$19.95

APA Member/Affiliate: \$19.95

ISBN 978-0-9792125-3-6

Item # 4441011



AD0535

To Order: 800-374-2721 • [www.apa.org/books](http://www.apa.org/books)



# BEST SELLERS

from the American Psychological Association



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

## Handbook of Counseling and Psychotherapy With Lesbian, Gay, Bisexual, and Transgender Clients

**Second Edition**  
*Edited by Kathleen J. Bieschke, Rupert M. Perez, and Kurt A. DeBord*  
2007. 464 pages. Hardcover.  
List: \$79.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-421-3 • Item # 4317113  
ISBN-13: 978-1-59147-421-0

## Career Paths in Psychology

**Where Your Degree Can Take You**  
**Second Edition**  
*Edited by Robert J. Sternberg*  
2007. 416 pages. Paperback.  
List: \$29.95  
APA Member/Affiliate: \$24.95  
ISBN 1-59147-732-8 • Item # 4313008  
ISBN-13: 978-1-59147-732-7

## Graduate Study in Psychology, 2007

2007. 832 pages. Paperback.  
List: \$24.95  
APA Member/Affiliate: \$21.95  
ISBN 1-59147-423-X • Item # 4270090  
ISBN-13: 978-1-59147-423-4

## Shy Children, Phobic Adults

**Nature and Treatment of Social Anxiety Disorder**  
**Second Edition**  
*Deborah C. Beidel and Samuel M. Turner*  
2007. 408 pages. Hardcover.  
List: \$59.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-452-3 • Item # 4317118  
ISBN-13: 978-1-59147-452-4

## Educating the Human Brain

*Michael I. Posner and Mary K. Rothbart*  
2007. 272 pages. Hardcover.  
List: \$79.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-381-0 • Item # 4318029  
ISBN-13: 978-1-59147-381-7

## Sensation Seeking and Risky Behavior

*Marvin Zuckerman*  
2007. 320 pages. Hardcover.  
List: \$59.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-738-7 • Item # 4317124  
ISBN-13: 978-1-59147-738-9

## Psychology and Economic Injustice

**Personal, Professional, and Political Intersections**  
*Bernice Lott and Heather E. Bullock*  
2007. 192 pages. Hardcover.  
List: \$59.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-429-9 • Item # 4316082  
ISBN-13: 978-1-59147-429-6

## Primate Perspectives on Behavior and Cognition

*Edited by David A. Washburn*  
2007. 368 pages. Hardcover.  
List: \$79.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-422-1 • Item # 4318035  
ISBN-13: 978-1-59147-422-7

## Becoming Culturally Oriented

**Practical Advice for Psychologists and Educators**  
*Nadya A. Fouad and Patricia Arredondo*  
2007. 208 pages. Hardcover.  
List: \$49.95  
APA Member/Affiliate: \$39.95  
ISBN 1-59147-424-8 • Item # 4317114  
ISBN-13: 978-1-59147-424-1

## Rumor Psychology

**Social and Organizational Approaches**  
*Nicholas DiFonzo and Prashant Bordia*  
2007. 392 pages. Hardcover.  
List: \$69.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-426-4 • Item # 4316079  
ISBN-13: 978-1-59147-426-5

## Scientific Jury Selection

*Joel D. Lieberman and Bruce D. Sales*  
2007. 264 pages. Hardcover.  
List: \$79.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-427-2 • Item # 4316081  
ISBN-13: 978-1-59147-427-2

## Why Aren't More Women in Science?

**Top Gender Researchers Debate the Evidence**  
*Edited by Stephen J. Ceci and Wendy M. Williams*  
2007. 248 pages. Hardcover.  
List: \$59.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-485-X • Item # 4316085  
ISBN-13: 978-1-59147-485-2

## Child Development and Social Policy

**Knowledge for Action**  
*Edited by J. Lawrence Aber, Sandra J. Bishop-Josef, Stephanie M. Jones, Kathryn Taaffe McLearn, and Deborah A. Phillips*  
2007. 352 pages. Hardcover.  
List: \$79.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-425-6 • Item # 4318036  
ISBN-13: 978-1-59147-425-8

## Preventing Youth Substance Abuse

**Science-Based Programs for Children and Adolescents**  
*Edited by Patrick Tolan, José Szapocznik, and Soledad Sambrano*  
2007. 264 pages. Hardcover.  
List: \$69.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-307-1 • Item # 4316058  
ISBN-13: 978-1-59147-307-7

## Contributions Toward Evidence-Based Psychocardiology

**A Systematic Review of the Literature**  
*Edited by Jacob Jordan, Benjamin Bardé, and Andreas Michael Zeiber*  
2007. 552 pages. Hardcover.  
List: \$99.95  
APA Member/Affiliate: \$59.95  
ISBN 1-59147-358-6 • Item # 4318028  
ISBN-13: 978-1-59147-358-9

## Dialogues on Difference

**Studies of Diversity in the Therapeutic Relationship**  
*Edited by J. Christopher Muran*  
2007. 336 pages. Hardcover.  
List: \$59.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-451-5 • Item # 4316083  
ISBN-13: 978-1-59147-451-7

## Disorders of the Self

**A Personality-Guided Approach**  
*Marshall L. Silverstein*  
2007. 320 pages. Hardcover.  
List: \$69.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-430-2 • Item # 4317116  
ISBN-13: 978-1-59147-430-2

## Intervening in Children's Lives

**An Ecological, Family-Centered Approach to Mental Health Care**  
*Thomas J. Dishion and Elizabeth A. Stormshak*  
2007. 320 pages. Hardcover.  
List: \$69.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-428-0 • Item # 4317115  
ISBN-13: 978-1-59147-428-9

## Insight in Psychotherapy

*Edited by Louis G. Castonguay and Clara E. Hill*  
2007. 488 pages. Hardcover.  
List: \$69.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-477-9 • Item # 4317122  
ISBN-13: 978-1-59147-477-7

## Second-Order Change in Psychotherapy

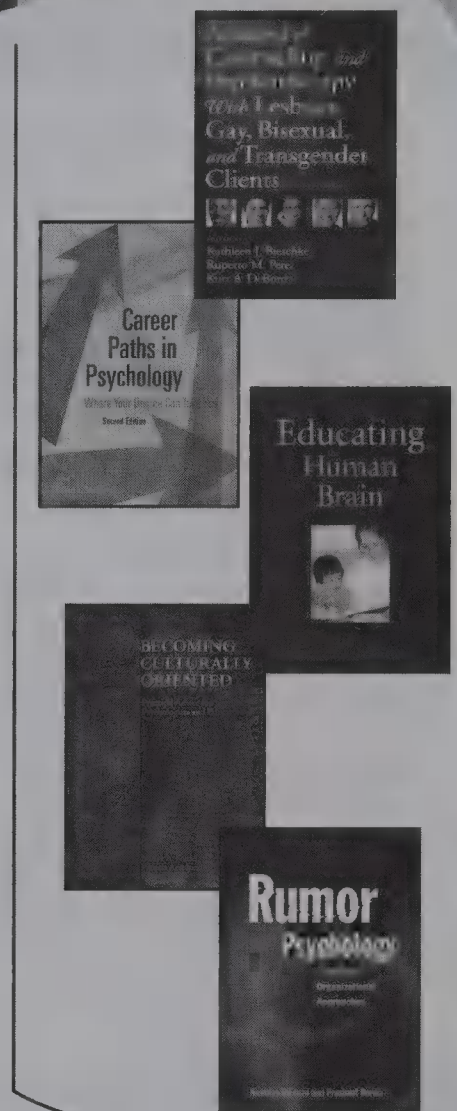
**The Golden Thread That Unifies Effective Treatments**  
*J. Scott Fraser and Andrew D. Solovey*  
2007. 312 pages. Hardcover.  
List: \$69.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-436-1 • Item # 4317117  
ISBN-13: 978-1-59147-436-4

## Psychological Treatment of Obsessive-Compulsive Disorder

**Fundamentals and Beyond**  
*Edited by Martin M. Antony, Christine Purdon, and Laura J. Summerfeldt*  
2007. 328 pages. Hardcover.  
List: \$59.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-484-1 • Item # 4317123  
ISBN-13: 978-1-59147-484-5

## Termination in Psychotherapy

**A Psychodynamic Model of Processes and Outcomes**  
*Anthony S. Joyce, William E. Piper, John S. Ogrodniczuk, and Robert H. Klein*  
2007. 224 pages. Hardcover.  
List: \$59.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-703-1 • Item # 4316086  
ISBN-13: 978-1-59147-703-3



## Psychology and the Department of Veterans Affairs

**A Historical Analysis of Training, Research, Practice, and Advocacy**  
*Rodney Baker and Wade Pickren*  
2007. 200 pages. Hardcover.  
List: \$59.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-453-1 • Item # 4316084  
ISBN-13: 978-1-59147-453-1

## Bereavement in Late Life

**Coping, Adaptation, and Developmental Influences**  
*Robert O. Hansson and Margaret S. Stroebe*  
2007. 232 pages. Hardcover.  
List: \$59.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-472-8 • Item # 4317119  
ISBN-13: 978-1-59147-472-2

## Spiritual Approaches in the Treatment of Women With Eating Disorders

*P. Scott Richards, Randy K. Hardman, and Michael E. Berrett*  
2007. 312 pages. Hardcover.  
List: \$59.95  
APA Member/Affiliate: \$49.95  
ISBN 1-59147-393-4 • Item # 4317103  
ISBN-13: 978-1-59147-393-0

To Order: 800-374-2721 • [www.apa.org/books](http://www.apa.org/books)



# BULLYING PREVENTION

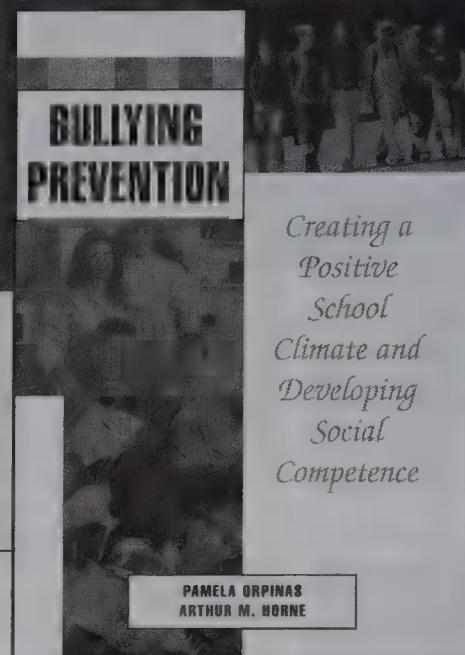
*Creating a Positive School Climate  
and Developing Social Competence*

PAMELA ORPINAS AND ARTHUR M. HORNE

**B**ullying is the most prevalent form of violence in American schools. In their new book *Bullying Prevention*, authors Pamela Orpinas and Andy Horne bring together years of experience in research and applied behavioral sciences to show how educators, school psychologists, counselors, and other professionals can address the problem of bullying and aggression in schools.

Readers will find definitions, statistics, and theories that will help them identify and characterize bullying. They will learn about the authors' School Social Development and Bullying Prevention Model, a blueprint for schools that students, teachers, and parents enjoy being a part of. This model shows how school professionals can prevent and reduce bullying by creating a positive environment and by ensuring all children have the social skills to communicate well and solve problems without aggression. The book has guidance in selecting research-based bullying prevention programs, and steps for assessing a school's needs and for evaluating a program's effectiveness.

The book also offers practical strategies for helping the children who are targets of bullying, and discusses counseling and family interventions for children who continue to bully despite positive changes in the school and classroom environment. A rich resource section contains a wide range of bullying-related readings, manuals, and tools available on the Internet. 2006. 320 pages. Hardcover.



## CONTENTS:

Introduction

### ■ PART I.

Understanding

the Problem ■ Chapter 1. Bullies: The Problem and the Impact

■ Chapter 2. Risks and Protective Factors for Bullying and Aggression

■ Chapter 3. Theoretical Perspectives on Bullying and Aggression

### ■ PART II. Addressing the Problem: Universal Interventions

■ Chapter 4. School Social Competence Development and Bullying

Prevention Model: The School ■ Chapter 5. School Social Competence

Development and Bullying Prevention Model: The Student ■ Chapter 6.

Evaluation of Bullying and Aggression Problems and Intervention

Programs ■ Chapter 7. Selection and Implementation of Universal

Bullying Prevention Programs ■ PART III. Addressing the Problem:

Persistent Bullies ■ Chapter 8. Persistent Bullying: Counseling

Interventions ■ Chapter 9. Persistent Bullying: Family Interventions

■ Chapter 10. Helping Children Who Are the Targets of Bullying

■ Resources ■ References

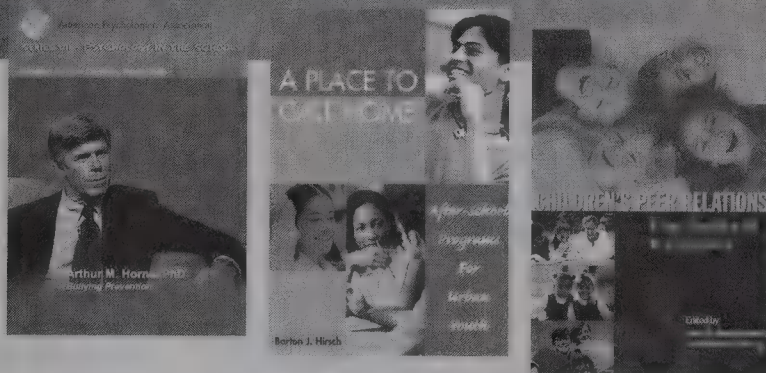
**List: \$59.95 • APA Member/Affiliate: \$49.95 • ISBN 1-59147-282-2 • Item # 4317082**

## ALSO AVAILABLE

### CHILDREN'S PEER RELATIONS From Development to Intervention

Edited by Janis B. Kupersmidt and Kenneth A. Dodge

2004 • 289 pages ■ Hardcover • List: \$59.95 ■ APA Member/  
Affiliate: \$44.95 ■ ISBN 1-59147-105-2 • Item # 4318008



**BULLYING PREVENTION** with Arthur M. Horne  
Part of the APA Psychotherapy Video Series VII:  
Psychology in the Schools

(VHS) ISBN 1-59147-340-3 ■ Item # 4310647

(DVD) ISBN 1-59147-341-1 ■ Item # 4310733

List: \$99.95 ■ APA Member/Affiliate: \$69.95



**A PLACE TO CALL HOME**  
After-School Programs for Urban Youth

Barton J. Hirsch

2005 • 163 pages ■ Hardcover • List: \$49.95 ■ APA Member/Affiliate: \$39.95

ISBN 1-59147-202-4 ■ Item # 4317059

**APA Books** Ordering Information

**800-374-2721**

**www.apa.org/books**

In Washington, DC, call: 202-336-5510

TDD/TTY: 202-336-6123 ■ Fax: 202-336-5502

**In Europe, Africa, or the Middle East,**

call: 44-207-240-0856



AMERICAN PSYCHOLOGICAL ASSOCIATION



# Child Development and Social Policy

## KNOWLEDGE FOR ACTION

Edited by J. Lawrence Aber, Sandra J. Bishop-Josef, Stephanie M. Jones, Kathryn Taaffe McLearn, and Deborah A. Phillips

Over the past 25 years, the intersection of developmental psychology and public policy has become an increasingly active and important domain for researchers, policymakers, children's rights advocates, and practitioners. At the forefront of the child development research and social policy movement is Edward Zigler, whose "knowledge for action" approach has revolutionized the way public policy is enacted to better serve vulnerable youth populations. *Child Development and Social Policy: Knowledge for Action* expands on Dr. Zigler's work in integrating the fields of child development and social policy, while using scientific knowledge for action as the model.

Contributors discuss these key questions: What are the most powerful research insights of the last 30 years that promote effective action for children and families? What are the most powerful constraints or limits of our knowledge base to promote effective action for children and families? What are the primary components of short-term research agenda to make the most powerful difference for children and families?

This edited volume focuses on both the influence of social policy on children's development and the unique perspective, insight, and skills that developmentalists bring to this policy and its formation. Programs to ensure good beginnings for all children are discussed, while the needs of those who are most vulnerable are also addressed. 2007. 352 pages. Hardcover.

Series: *Decade of Behavior*

List: \$79.95 • APA Member/Affiliate: \$49.95 • ISBN 1-59147-425-6 • Item # 4318036 • ISBN-13: 978-1-59147-425-8

### CONTENTS:

Acknowledgments ■ Introduction  
 ■ Part I: Making History: Child Development and Social Policy  
 ■ Chapter 1. Child Development Research and Public Policy: Triumphs and Setbacks on the Way to Maturity ■ Chapter 2. Policy Looking to Research ■ Chapter 3. Bridging the Gap Between Research and Child Policy Change: The Role of Strategic Communications in Policy Advocacy ■ Chapter 4. Data for a Democracy: The Evolving Role of Evaluation in Policy and Program Development ■ Part II: Ensuring Good Beginnings for All Children ■ Chapter 5. Forty Years of Research Knowledge and Use: From Head Start to Early Head Start and Beyond ■ Chapter 6. Beyond Baby Steps: Promoting the Growth and Development of U.S. Child Care Policy ■ Chapter 7. From Visions to Systems of Universal Pre-kindergarten ■ Chapter 8. Policies to Ensure that No Child Starts from Behind ■ Part III: Addressing the Needs of the Most Vulnerable Children and Families ■ Chapter 9. Poverty and Child Development: New Perspectives on a Defining Issue ■ Chapter 10. Intervention and Policy Implications of Research on Neurobiological Functioning in Maltreated Children ■ Chapter 11. The Sexually Mature Teen as a Whole Person: New Directions in Prevention and Intervention for Teen Pregnancy and Parenthood ■ Chapter 12. Children in Foster Care ■ Part IV: Strengthening Children, Families, and Communities ■ Chapter 13. Parent Education: Lessons Inspired by Head Start ■ Chapter 14. Mental Health: A Neglected Partner in the Healthy Development of Young Children ■ Chapter 15. Family Support: A Force for Change ■ Chapter 16. Using the Web to Disseminate Research and Affect Public Policy ■ Epilogue

## ALSO AVAILABLE

### CHILDREN'S PEER RELATIONS

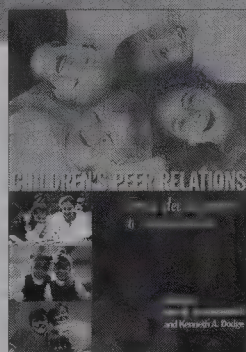
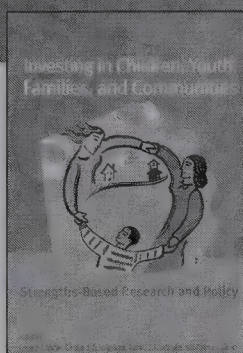
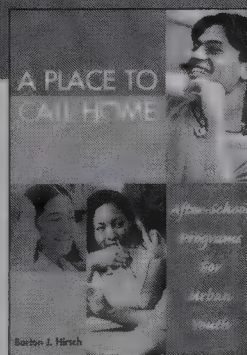
#### From Development to Intervention

Edited by Janis B. Kupersmidt and Kenneth A. Dodge

2004 ■ 289 pages ■ Hardcover ■ List: \$59.95

APA Member/Affiliate: \$44.95 ■ ISBN 1-59147-105-2

Item # 4318008 ■ ISBN-13: 978-1-59147-105-9



### A PLACE TO CALL HOME

#### After-School Programs for Urban Youth

Barton J. Hirsch

2005 ■ 163 pages ■ Hardcover ■ List: \$49.95 ■ APA Member/Affiliate: \$39.95

ISBN 1-59147-202-4 ■ Item # 4317059 ■ ISBN-13: 978-1-59147-202-5

### INVESTING IN CHILDREN, YOUTH, FAMILIES, AND COMMUNITIES: Strengths-Based Research and Policy

Edited by Kenneth I. Maton, Cynthia J. Schellenbach, Bonnie J. Leadbeater, and Andrea L. Solarz

2004 ■ 380 pages ■ Hardcover ■ List: \$49.95 ■ APA Member/Affiliate: \$39.95

ISBN 1-59147-062-5 Item # 4316021 ■ ISBN-13: 978-1-59147-062-5

## APA Books Ordering Information

**800-374-2721**

**www.apa.org/books**

In Washington, DC, call: 202-336-5510

TDD/TTY: 202-336-6123 ■ Fax: 202-336-5502

In Europe, Africa, or the Middle East,

call: 44-207-240-0856



AMERICAN PSYCHOLOGICAL ASSOCIATION



# NEW RELEASES

from the American Psychological Association

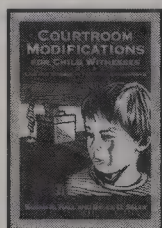


AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION



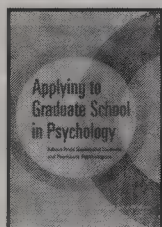
## Handbook of Bereavement Research and Practice

Advances in Theory and Intervention  
*Edited by Margaret S. Stroebe, Robert O. Hansson, Henk Schut, and Wolfgang Stroebe*  
2008. 624 pages. Hardcover.  
ISBN 978-1-4338-0351-2 ■ Item # 4318045  
List: \$69.95 ■ APA Member/Affiliate: \$49.95



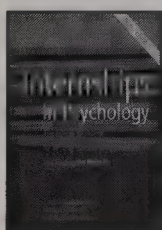
## Courtroom Modifications for Child Witnesses

Law and Science in Forensic Evaluations  
*Susan R. Hall and Bruce D. Sales*  
2008. 368 pages. Hardcover.  
ISBN 978-1-4338-0354-3 ■ Item # 4317156  
List: \$79.95 ■ APA Member/Affiliate: \$49.95



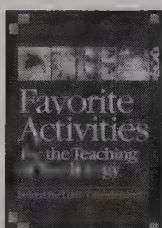
## Applying to Graduate School in Psychology

Advice From Successful Students and Prominent Psychologists  
*Edited by Amanda C. Kracen and Ian J. Wallace*  
2008. 304 pages. Paperback.  
ISBN 978-1-4338-0345-1 ■ Item # 4313018  
List: \$34.95 ■ APA Member/Affiliate: \$29.95



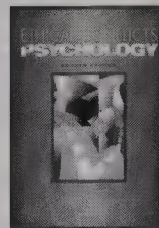
## Internships in Psychology

The APAGS Workbook for Writing Successful Applications and Finding the Right Fit  
Second Edition  
*Carol Williams-Nickelson, Mitchell J. Prinstein, and W. Gregory Keilin*  
2008. 144 pages. Paperback.  
ISBN 978-1-4338-0355-0 ■ Item # 4313021  
List: \$24.95 ■ APA Member/Affiliate: \$19.95



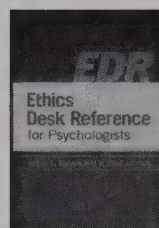
## Favorite Activities for the Teaching of Psychology

*Edited by Ludy T. Benjamin, Jr.*  
2008. 400 pages. Paperback.  
ISBN 978-1-4338-0349-9 ■ Item # 4316105  
List: \$34.95 ■ APA Member/Affiliate: \$29.95



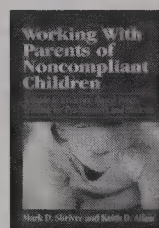
## Ethical Conflicts in Psychology

Fourth Edition  
*Donald N. Bersoff*  
2008. 632 pages.  
Hardcover.  
ISBN 978-1-4338-0350-5 ■ Item # 4312009  
List: \$69.95 ■ APA Member/Affiliate: \$54.95  
Paperback.  
ISBN 978-1-4338-0353-6 ■ Item # 4312012  
List: \$49.95 ■ APA Member/Affiliate: \$39.95



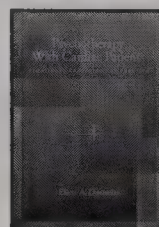
## Ethics Desk Reference for Psychologists

*Jeffrey E. Barnett and W. Brad Johnson*  
2008. 200 pages. Spiral-bound.  
ISBN 978-1-4338-0352-9 ■ Item # 4312011  
List: \$39.95 ■ APA Member/Affiliate: \$34.95



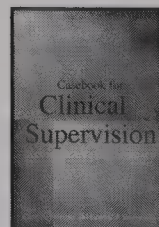
## Working With Parents of Noncompliant Children

A Guide to Evidence-Based Parent Training for Practitioners and Students  
*Mark D. Shriver and Keith D. Allen*  
2008. 280 pages. Hardcover.  
ISBN 978-1-4338-0344-4 ■ Item # 4317155  
List: \$59.95 ■ APA Member/Affiliate: \$49.95



## Psychotherapy With Cardiac Patients

Behavioral Cardiology in Practice  
*Ellen A. Dornelas*  
2008. 280 pages. Hardcover.  
ISBN 978-1-4338-0356-7 ■ Item # 4317157  
List: \$59.95 ■ APA Member/Affiliate: \$49.95

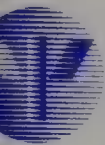


## Casebook for Clinical Supervision

A Competency-Based Approach  
*Edited by Carol A. Falender and Edward P. Shafranske*  
2008. 272 pages. Hardcover.  
ISBN 978-1-4338-0342-0 ■ Item # 4317154  
List: \$59.95 ■ APA Member/Affiliate: \$49.95

AD0588

To Order: 800-374-2721 • [www.apa.org/books](http://www.apa.org/books)



Volume 100  
Number 3

August 2008

Published quarterly  
by the  
American Psychological  
Association

ISSN 0022-0663

# Journal of Educational Psychology

Margaret R. Harris, *Editor*

Eric M. Anderman, *Associate Editor*

Donna M. Kulikowich, *Associate Editor*

Victoria Miller, *Associate Editor*

Frank Pajares, *Associate Editor*

Jeffrey J. Walczyk, *Associate Editor*

CURRENT YR/VOL

Marygrove College Library  
8425 West McNichols Road  
Detroit, MI 48221

[www.apa.org/journals/edu](http://www.apa.org/journals/edu)

2000-2010  
DECADE  
of BEHAVIOR



- 672 An Exploration of Young Adolescents' Social Achievement Goals and Social Adjustment in Middle School  
*Allison M. Ryan and S. Serena Shim*
- 688 Students' Motivational Profiles and Achievement Outcomes in Physical Education: A Self-Determination Perspective  
*Julie C. S. Boiché, Philippe G. Sarrazin, Frederick M. E. Grouzet, Luc G. Pelletier, and Julien P. Chanal*
- 702 Teachers' Occupational Well-Being and Quality of Instruction: The Important Role of Self-Regulatory Patterns  
*Uta Klusmann, Mareike Kunter, Ulrich Trautwein, Oliver Lüdtke, and Jürgen Baumert*
- 716 Pedagogical Content Knowledge and Content Knowledge of Secondary Mathematics Teachers  
*Stefan Krauss, Martin Brunner, Mareike Kunter, Jürgen Baumert, Werner Blum, Michael Neubrand, and Alexander Jordan*

## Other

- 654 American Psychological Association Subscription Claims Information
- 602 E-Mail Notification of Your Latest Issue Online!
- 726 Instructions to Authors
- 565 Low Publication Prices for APA Members and Affiliates
- 509 Members of Underrepresented Groups: Reviewers for Journal Manuscripts Wanted
- ii Subscription Order Form

## ORDER FORM

Start my 2008 subscription to the *Journal of Educational Psychology* ISSN: 0022-0663

\_\_\_\_\_ \$73.00, APA MEMBER/AFFILIATE \_\_\_\_\_

\_\_\_\_\_ \$161.00, INDIVIDUAL NONMEMBER \_\_\_\_\_

\_\_\_\_\_ \$450.00, INSTITUTION \_\_\_\_\_

*In DC add 5.75% / In MD add 6% sales tax*

**TOTAL AMOUNT ENCLOSED** \$ \_\_\_\_\_

**Subscription orders must be prepaid.** (Subscriptions are on a calendar year basis only.) Allow 4-6 weeks for delivery of the first issue. Call for international subscription rates.



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

**SEND THIS ORDER FORM TO:**  
American Psychological Association  
Subscriptions  
750 First Street, NE  
Washington, DC 20002-4242

Or call **800-374-2721**, fax **202-336-5568**.  
TDD/TTY **202-336-6123**.  
For subscription information, e-mail:  
**subscriptions@apa.org**

☐ **Check enclosed** (make payable to APA)

**Charge my:** ☐ VISA ☐ MasterCard ☐ American Express

Cardholder Name \_\_\_\_\_

Card No. \_\_\_\_\_ Exp. Date \_\_\_\_\_

\_\_\_\_\_  
Signature (Required for Charge)

### BILLING ADDRESS:

Street \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

Daytime Phone \_\_\_\_\_

E-mail \_\_\_\_\_

### MAIL TO:

Name \_\_\_\_\_

Address \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

APA Member # \_\_\_\_\_ EDUA08

# Effects of Small-Group Tutoring With and Without Validated Classroom Instruction on At-Risk Students' Math Problem Solving: Are Two Tiers of Prevention Better Than One?

Lynn S. Fuchs, Douglas Fuchs, Caitlin Craddock,  
Kurstin N. Hollenbeck, and Carol L. Hamlett  
Vanderbilt University

Christopher Schatschneider  
Florida State University

This study assessed the effects of small-group tutoring with and without validated classroom instruction on at-risk students' math problem solving. Stratifying within schools, 119 3rd-grade classes were randomly assigned to conventional or validated problem-solving instruction (Hot Math, schema-broadening instruction). Students identified as at risk ( $n = 243$ ) were randomly assigned, within classroom conditions, to receive or not receive Hot Math tutoring. Students were tested on problem-solving and math applications measures before and after 16 weeks of intervention. Analyses of variance, which accounted for the nested structure of the data, revealed that the tutored students who received validated classroom instruction achieved better than the tutored students who received conventional classroom instruction (effect size = 1.34). However, the advantage for tutoring over no tutoring was similar whether students received validated or conventional classroom instruction (effect sizes = 1.18 and 1.13). Tutoring, not validated classroom instruction, reduced the prevalence of math difficulty. Implications for responsiveness-to-intervention prevention models and for enhancing math problem-solving instruction are discussed.

**Keywords:** mathematics, mathematics instruction, mathematics disability

**Supplemental materials:** <http://dx.doi.org/10.1037/0022-0063.100.3.491.supp>

Mathematics word problems require the transfer of knowledge to novel situations, and this form of transfer can be difficult to effect (cf. Bransford & Schwartz, 1999; Mayer, Quilici, & Moreno, 1999), especially for primary-grade children (Durnin, Perrone, & MacKay, 1997; Foxman, Ruddock, McCallum, & Schagen, 1991, cited in Boaler, 1993; Larkin, 1989). Whereas a calculations problem is already set up for solution, a word problem requires students to use text to determine what information is unknown and to construct and solve a number sentence for finding that unknown information. So, although calculation skill is foundational to word problems (e.g., L. S. Fuchs et al., 2006), it seems likely that for some students, word-problem skill will not develop automatically from instruction on math calculations, but rather that instruction needs to be designed specifically to prevent difficulty with word problems.

---

Lynn S. Fuchs, Douglas Fuchs, Caitlin Craddock, Kurstin N. Hollenbeck, and Carol L. Hamlett, Department of Special Education, Vanderbilt University; Christopher Schatschneider, Department of Psychology, Florida State University.

This research was supported in part by Grant 1 RO1 HD46154 and by Core Grant HD15052 from the National Institute of Child Health and Human Development to Vanderbilt University. Statements do not reflect the position or policy of these agencies, and no official endorsement by them should be inferred.

Correspondence concerning this article should be addressed to Lynn S. Fuchs, 228 Peabody, Vanderbilt University, Nashville, TN 37203. E-mail: [lynn.fuchs@vanderbilt.edu](mailto:lynn.fuchs@vanderbilt.edu)

The context for preventing academic difficulty in the schools has changed over the past 5 years with the introduction of multi-tiered prevention systems.<sup>1</sup> Adapted from the health care system, school-based multi-tier prevention systems typically involve three tiers. The first tier is research-principled or validated classroom instruction. Students who are deemed at risk for difficulty with the classroom program (AR students), usually on the basis of screening near the beginning of the school year, also receive a second tier of prevention, with a standard, validated small-group tutoring protocol (that can be expected to benefit most students). Only students who prove unresponsive to classroom instruction and to tutoring are referred for a comprehensive evaluation to consider the possibility of a disability that requires a third, more individualized tier of prevention, usually special education. Because such a multitier prevention system involves assessing a student's responsiveness-to-intervention (RTI), it is conventionally referred to as an *RTI prevention system* (e.g., see Vaughn & Fuchs, 2003).

Most RTI studies have addressed reading (e.g., D. Fuchs, Compton, Fuchs, & Davis, in press; Vaughn, Linan-Thompson, & Hickman, 2003; Vellutino et al., 1996). In mathematics, where fewer studies have been conducted, the bulk of the literature has been conducted by VanDerHeyden, Witt, and their colleagues. For example, VanDerHeyden, Witt, and Gilbertson (2007) screened

---

<sup>1</sup> In the RTI literature, any instructional program that is grade-level appropriate is conceptualized as a form of prevention, not remediation. In the present study, instruction focused on third-grade curricular targets, making the use of the term *prevention* appropriate.



elementary-grade students on grade-level computational fluency. Classrooms scoring below a given criterion received classroom instruction to build computational fluency. Individual students who scored below the 16th percentile of their own class, with classroom intervention when needed, were assessed to determine whether students could score above the 16th percentile if offered a reward. Only students who could not achieve the criterion even when rewarded entered tutoring, which involved 10 min per day of computation instruction with modeling, guided practice with immediate error correction, independent timed practice with slightly delayed error correction, and the opportunity to earn a reward. VanDerHeyden et al. evaluated this multi-tiered RTI math prevention model, using a multiple-baseline across-schools design with five schools. Referrals to and evaluations for special education decreased with implementation, but effects on students' math performance were not reported. In related work, VanDerHeyden and Witt (2005) used a similar model to screen first and second graders in one school, using basic fact assessments. With the subset of children who scored poorly, nine practice sessions with feedback and reinforcement were conducted to increase basic fact fluency. On the basis of these sessions, the researchers identified nonresponders and examined decision utility with respect to final outcomes, but they did not explore effects on student learning.

These studies and others in VanDerHeyden's line of work are ambitious and important, but they are restricted to calculations, rely on a narrow approach to math instruction, and do not assess effects on students' math performance. L. S. Fuchs et al. (2005) extended this line of work by conducting a randomized field trial to assess the effects of second-tier small-group tutoring on a broader first-grade math curriculum, including number sense and concepts, addition and subtraction concepts and procedures, number combinations, missing addends, place value, and word problems. A more comprehensive instructional design and a more sustained intervention were used, and the study supported the efficacy of the small-group tutoring on students' learning outcomes. However, the focus on word problems was limited to 1 week of the 16-week tutoring program, and the researchers did not consider whether tutoring efficacy was enhanced when classroom instruction relied on a validated approach.

In fact, we identified only one prior RTI or prevention study that assessed the efficacy of tutoring with and without validated classroom instruction. It was conducted with first graders and assessed reading tutoring. Gilbert, Compton, Fuchs, Fuchs, and Schatschneider (2007) showed that tutoring was similarly effective for AR students with or without validated classroom reading instruction as the backdrop for tutoring. However, in contrast to most RTI models where risk status is identified on the basis of initial screening, AR students entered Gilbert et al.'s study only after a full semester of progress monitoring showed inadequate response to the classroom instructional program; thus, it is not surprising that the nature of classroom instruction during tutoring did not enhance tutoring's efficacy.

The issue of how tutoring interacts with classroom instruction among AR students who may have poor learning outcomes is important for designing efficient and effective RTI prevention systems. If tutoring is differentially efficacious when combined with validated classroom instruction, then both tiers are critical, and classroom instruction needs to be designed deliberately with AR students in mind, even when they receive tutoring. By contrast,

if tutoring promotes comparable outcomes regardless of the classroom instructional context, then findings would set the stage for research on whether tutoring might occur as a replacement for, rather than as a supplement to, classroom instruction. Tutoring as a replacement for classroom instruction would make RTI prevention systems more feasible and efficient and would permit resources to be infused at the tutoring tier. Of course, in considering tutoring efficacy as a function of the classroom's instructional context, it is important not only to compare the response of AR students who do and do not receive tutoring, but also to contrast the learning of those AR students against not-at-risk (NAR) students. This has not been addressed in prior work.

The purpose of the present study, therefore, was to examine students' mathematics problem-solving learning and to explore the prevalence of mathematics difficulty as a function of validated classroom prevention, as a function of small-group tutoring, and as a function of whether tutoring occurs with or without validated classroom instruction. (We refer to *small-group tutoring* as *tutoring* in this article.) We describe the theoretical basis for our approach to word-problem instruction and summarize previous related work. Then we consider previous studies assessing the effects of one or two tiers of prevention on the prevalence of mathematics difficulty. Finally, we review the study's purpose and design.

### Transfer, Schema Theory, and Related Work on Word-Problem Instruction

The mathematics education literature shows that children experience difficulty transferring the math competence they develop in school. For example, Larkin (1989) demonstrated that children fail to apply simple computational skills when problems change in minor ways. Other work (Foxman, Ruddock, McCallum, & Schagen, 1991, cited in Boaler, 1993) has illustrated how students fail to realize connections between math problems presented in and out of context. Such findings are supported by theoretical frameworks for understanding the development of transfer that challenge the assumption of vertical transfer, whereby mastery of simple skills facilitates acquisition of more complex skills (Gagne, 1968; Resnick & Resnick, 1992). The notion of vertical transfer has been replaced with the concept of lateral transfer, by which children recognize problems across numerous experiences to abstract generalized problem-solving strategies.

Some refer to the abstraction of generalized problem-solving strategies as the *development of schemas* (A. L. Brown, Campione, Webber, & McGilly, 1992; Gick & Holyoak, 1983). A schema is a category that encompasses similar problems; it is a problem type (Chi, Feltovich, & Glaser, 1981; Gick & Holyoake, 1983; Mayer, 1992; Quilici & Mayer, 1996). The broader the category for the problem type, the greater the probability that individuals will recognize connections between familiar and novel problems so that they will know when to apply the solution methods they have learned. Essentially, students expand the domain of problems for which they recognize that they have the mathematical means to find answers, thereby facilitating transfer to novel problems. To facilitate schema development, teachers must first teach problem-solution rules, as they conventionally do. Then, however, teachers must help students develop schemas and awareness of those schemas (Cooper & Sweller, 1987) and must provide students with a wealth of problem-solving activities that emphasize knowledge



application (J. S. Brown, Collins, & Duguid, 1989; Prawat, 1992). The hope is that schemas can influence behavior in a broad set of situations and thereby affect breadth of learning or transfer (A. L. Brown et al., 1992; Glaser, 1983). Research has substantiated the importance of mastering rules for problem solution (e.g., Mawer & Sweller, 1985; Sweller & Cooper, 1985), but less is known about how to help students develop schemas and awareness of those schemas (e.g., Bransford & Schwartz, 1999; Cooper & Sweller, 1987; Mayer et al., 1999).

Some studies have relied on explicit instruction based on schema theory to enhance word-problem skill. Working primarily at the second or third tier of the RTI prevention system in small groups, Jitendra and colleagues have demonstrated acquisition, maintenance, and transfer effects for students with serious math deficits or with risk for math difficulty at eighth grade (Jitendra, DiPipi, & Perron-Jones, 2002), sixth grade (Xin, Jitendra, & Deatline-Buchman, 2005), and third and fourth grades (Jitendra et al., 1998, 2007; Jitendra & Hoff, 1996). In our work, we have also relied on schema theory. In contrast to the studies of Jitendra and colleagues, our work has been conducted exclusively at third grade, with more difficult problem types, with a focus on far transfer that extends beyond the transfer requirements in Jitendra's studies, and with a key difference in the way schema theory is applied. In instruction similar to Jitendra's schema-based strategy instruction, we have taught students to understand the underlying mathematical structure of the problem type, to recognize the basic problem-type category, and to solve the problem type. In contrast to Jitendra's teaching methods, we have incorporated a fourth instructional feature by explicitly teaching students to transfer their problem-solving skills. In keeping with Cooper and Sweller (1987), our goal has been to help students recognize connections between problems like those worked during instruction and problems with unexpected features, such as problems that include irrelevant information, that present a novel question requiring an extra step, that include relevant information presented in charts or graphs, or that combine problem types, and so on. Our hope has been that the addition of explicit transfer instruction would lead to more flexible and successful problem solving. We refer to the combination of all four instructional components as *schema-broadening instruction* (SBI). Teachers and students refer to this approach as *Hot Math*.

In our first randomized field trial, L. S. Fuchs, Fuchs, et al. (2003b) separated the effects of the first three instructional components (teaching students to understand the underlying mathematical structure of the problem type, to recognize the basic schema for a problem type, and to solve the problem type) from those of the last instructional phase (explicitly teaching students to transfer). The study was conducted at the first tier of the RTI prevention system, in general education classrooms, with whole-class instruction, and with all students for whom we had consent. The basic word-problem types targeted for instruction were more complex than had yet been studied with the third graders, including finding half, step-up functions, two-step problems involving pictographs, and two- and three-step shopping list problems. Also, far transfer was assessed on taught and untaught problem types within a highly novel and complex context that resembled real-life problem-solving situations. Third-grade classes were randomly assigned to teacher-designed word-problem instruction, to experimenter-designed instruction on the first three instructional

components, or to experimenter-designed SBI that incorporated the first three components but also explicitly taught students to transfer. With the addition of this last instructional component, teachers explained how problem features, such as format or vocabulary, can make problems seem unfamiliar without modifying the problem type or the required solution rules; they discussed examples, emphasizing that they had the same problem types as superficial features, such as format or vocabulary, changed; they provided practice in sorting novel problems in terms of changes in superficial problem features and in terms of solving those problems; and they prompted students to search novel problems for familiar problem types. Results indicated that on near- and far-transfer word-problem tasks, Hot Math SBI (i.e., all four components) strengthened problem-solving performance over and beyond experimenter-designed instruction that addressed only the first three components (understanding the underlying mathematical structure, recognizing the basic schema, and solving problems) and beyond teacher-designed instruction.

In subsequent work (L. S. Fuchs, Fuchs, Finelli, et al., 2004), SBI that addressed transfer instruction on three superficial features was more effective than SBI that addressed transfer instruction on six superficial features. However, the present study is the first to assess the efficacy of SBI tutoring on relatively complex problem types. Moreover, in our own prior work, as in other math and reading RTI research, the combined and separate effects of classroom instruction and tutoring have not been assessed, and the prevalence of math difficulty with one and/or two tiers of prevention has not been examined.

### Prevalence of Math Difficulty

Studies (e.g., Gross-Tsur, Manor, & Shalev, 1996; Lewis, Hitch, & Walker, 1994; Shalev, 2007) have indicated that 5%–9% of the school-age population experience some form of math disability. Four problems, however, limit the utility of available figures. First, much of the literature is limited to math facts and simple computation. Second, many studies do not rely on individual tests, as used in the schools to identify disability. Third, to recruit adequately large samples, high cut scores have often been applied. For example, Jordan and Hanich (2000) and Jordan and Montani (1997) demarcated math difficulty as falling below the 30th percentile on a group-administered test; Hanich, Jordan, Kaplan, and Dick (2001) used the 35th percentile. Russell and Ginsburg (1984) included students performing one level below their expected grade, and Geary (1990) identified children participating in Chapter 1 (in which mean study scores corresponded to percentile ranks of 25 in Grade 1 and 40 in Grade 2). Fourth, in the absence of sound prevention activities, math disability estimates may be artificially high because they fail to eliminate inadequate instruction as a possible explanation for low math achievement.

In the present study, we sought to address these problems. We focused on applied math problems. We investigated the prevalence of math difficulty on an individually administered math test as is done in the schools to identify disability, and we applied a cut score more similar to those used in the schools (i.e., 15th percentile). Finally, to understand how sound prevention activities may affect the prevalence of math disability, we examined disability prevalence with and without SBI classroom instruction and with and without SBI tutoring.



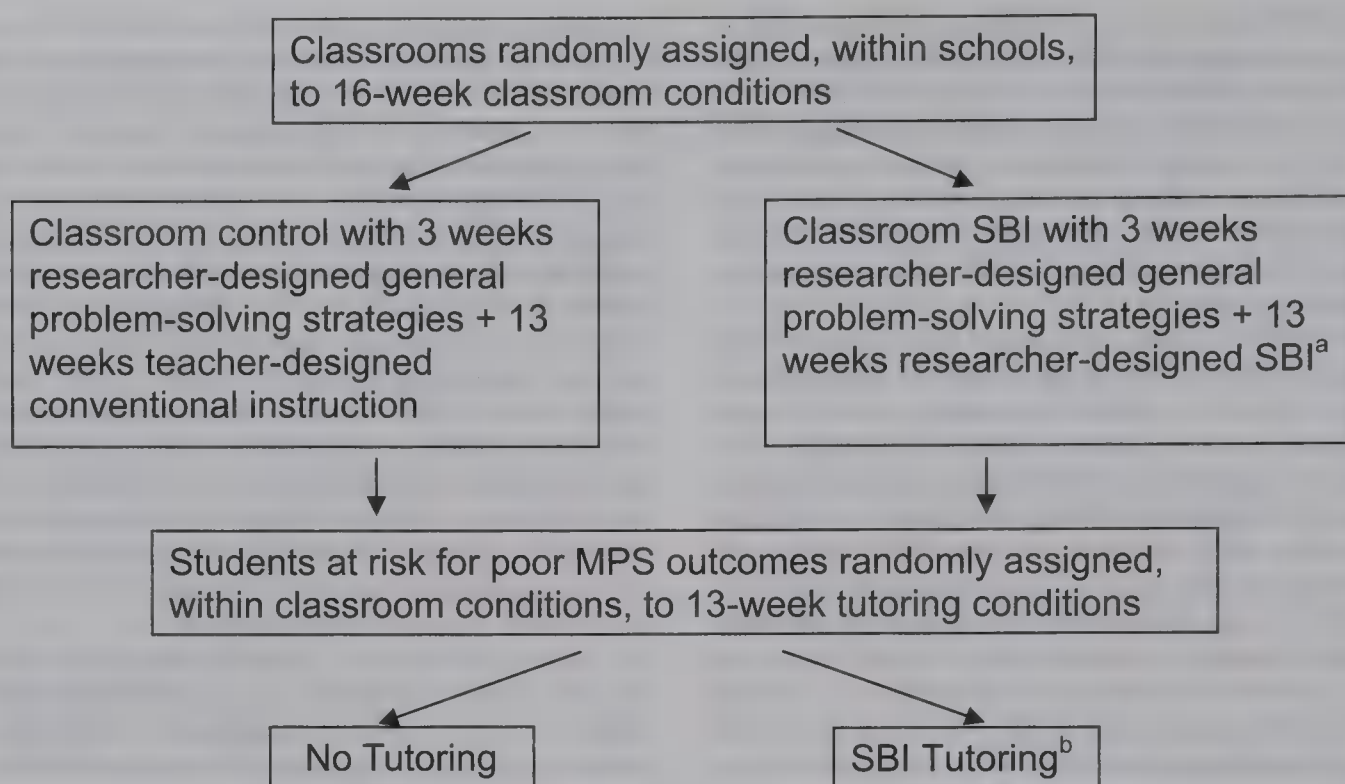


Figure 1. The study design is shown. SBI = schema-broadening instruction; MPS = math problem solving.

<sup>a</sup>In Years 1–3, SBI classrooms were randomly assigned to receive standard SBI or a variant of SBI. <sup>b</sup>In Years 1–2, SBI tutoring students were randomly assigned to receive standard SBI tutoring or a variant of SBI tutoring

### Study Purposes and Overview

The primary purpose of the present study was to assess the effects of SBI tutoring, with and without SBI classroom instruction, on AR students' math problem solving and on their learning relative to NAR peers. We also examined the prevalence of mathematics difficulty with and without one or two tiers of SBI math problem-solving prevention, and we assessed the efficacy of the classroom and tutoring SBI instructional protocols. Within schools, we randomly assigned third-grade classrooms to receive classroom SBI or to receive the math problem-solving instruction designed by their teachers. In these classrooms, we designated NAR and AR students on the basis of their incoming math scores and, within classroom conditions, randomly assigned AR students to continue in their classroom conditions with or without SBI tutoring. Students were pre- and posttested in a 3-week window before and after prevention occurred (see Figure 1 for the study design).

### Method

#### Participants

**Classrooms and their teachers.** In a southeastern metropolitan school district, 120 third-grade classrooms participated in this study. Stratifying so that each condition was represented approximately comparably in each school, we randomly assigned 40 classrooms to the control condition (i.e., 3 weeks of researcher-designed general math problem-solving instruction plus 13 weeks of teacher-designed math problem-solving instruction) and assigned 80 classrooms to the Hot Math SBI condition (i.e., 3 weeks of researcher-designed general math problem solving plus 13

weeks of researcher-designed SBI). The study occurred over 4 school years during a time when the school district was relatively stable.<sup>2</sup> One quarter of the sample entered the study each year. During the first 3 years, SBI classrooms were randomly assigned to Hot Math SBI or to a variant designed to strengthen Hot Math SBI (see Table 1 for a summary of the classroom SBI variants). In the first three cohorts, the effects of the two SBI conditions were not statistically significantly different, but both were reliably better than control. Therefore, we did not test a variant in Cohort 4 (all Cohort 4 teachers were randomly assigned to control or to the standard version of SBI). We considered all students in the SBI classrooms to have participated in one SBI condition; however, to assess SBI variants, we included cohort effects in the analytic model. One Cohort 3 classroom in the SBI condition left the study during the first month of participation because of the classroom teacher's personal reasons. The remaining 119 classroom teachers

<sup>2</sup> Across the 4 years of the study, the school district changed as follows. The student population increased from 69,445 to 70,140. The number of schools increased from 126 to 133. The number of teachers decreased from 4,831 to 4,809. The number of White students decreased from 31,460 to 28,483, whereas the number of Hispanics rose from 7,017 to 10,467, and the number of African Americans increased from 35,349 to 36,864. Students classified as having limited English proficiency increased by 1.9%; those classified with disabilities decreased by 1.6%. Per pupil spending increased from \$8,885 to \$9,300. In 2005–2006, a new math core instructional program was introduced (see description of classroom control condition). In terms of administration, the only significant change was the departure of the chief instructional officer. Because this occurred during the last month of implementation of the 4-year study, it did not affect instructional context.

Table 1  
Instructional Variants by Tier and by Cohort

Cohort	Variant	Incorporated as standard SBI in subsequent years
Classroom tier		
1	Once per week, RA teacher led discussion of how transfer features apply to a video portraying a real-life math situation that included second-grade problem types	No
2	In each session, in paired practice, students extended one partially developed problem into a specific problem type. This was guided by problem production steps. (What kind of problem? What do you know? What should you add?)	No
3	For every problem, students asked themselves meta-questions to guide the problem-solving process. (Who is the most important person in the story? What is the most important thing the who is doing? What question do we need to answer with math? What's the problem type? What transfer features are included?)	No
Tutoring tier <sup>a</sup>		
1	Self-regulated learning strategies incorporated within tutoring sessions	Yes and strengthened in Cohorts 2, 3, and 4 to enhance motivation
2	Drill-and-practice warm-ups to build fluency with a foundational skill necessary for the specific problem type	No

*Note.* When variants required additional time, students in the standard SBI condition had more practice to keep instructional time comparable. SBI = schema-broadening instruction; RA = research assistant.

<sup>a</sup>See text for more information.

were predominantly female (97.4% in control, 97.5% in SBI) and European American (79.5% in control, 66.7% in SBI; the teachers who were not European American were African American). In the control group, teachers had taught for an average of 14.18 years ( $SD = 10.49$ ), and in the SBI group, teachers had taught for an average of 16.09 years ( $SD = 10.27$ ). Average class size for the control group was 18.69 ( $SD = 2.40$ ), and average class size for the SBI group was 18.45 ( $SD = 2.62$ ). There were no significant differences on these variables by classroom condition.

Research assistants, not classroom teachers, delivered researcher-designed general problem-solving strategies instruction to all 119 classrooms during the first 3 weeks of instruction. For the final 13 weeks of instruction, research assistants delivered classroom SBI, whereas classroom teachers delivered conventional instruction. We refer to research assistants who taught general problem-solving strategies or classroom SBI as *RA teachers*, and we refer to the real classroom teachers as *classroom teachers*. Whenever an RA teacher was delivering instruction, the classroom teacher was present to help distribute materials, maintain order, and provide corrective feedback. (When there was more than one variant of classroom SBI, each RA teacher taught in both conditions.) We also note that the math program in all classrooms went beyond the instructional procedures described in the study; that is, experimental procedures were incorporated within the classroom teachers' broader mathematics instructional program. We describe research assistants later.

*Students.* To obtain a representative sample, we screened 2,023 students for whom we had consent. That is, in the 119 third-grade classrooms, we randomly sampled 1,200 students for participation, blocking within classrooms and within three strata: (a) 25% of students with scores 1 standard deviation below the mean of the entire distribution on the Test of Computational

Fluency (L. S. Fuchs, Hamlett, & Fuchs, 1990); (b) 50% of students with scores within 1  $SD$  of the mean of the entire distribution on the Test of Computational Fluency; and (c) 25% of students with scores 1 standard deviation above the mean of the entire distribution on the Test of Computational Fluency. Of these 1,200 students, 59 moved prior to posttesting (including 45 AR students; see later discussion). The 59 children who moved prior to posttesting were demographically comparable to the pupils who remained in the study. Among the remaining 1,141 students, IQ, as measured on the two-subset Wechsler Abbreviated Intelligence Scale (Wechsler, 1999), averaged 97.29 ( $SD = 13.93$ ). Standard scores on the Woodcock-Johnson III Tests of Achievement (WJ III; Woodcock, McGrew, & Mather, 2001) Applied Problems averaged 102.56 ( $SD = 13.60$ ), and standard scores on the Woodcock Reading Mastery Test-Revised (Woodcock, 1998) Word Identification averaged 100.83 ( $SD = 10.09$ ). Of these students, 559 (49.0%) were male, and 626 (54.9%) received subsidized lunch. Ethnicity was distributed as follows: 480 (42.1%) African American, 464 (40.7%) European American, 120 (10.5%) Hispanic, 17 (1.5%) Kurdish, and 60 (5.3%) other. Thirty-seven students (3.2%) were English language learners.

Among these students, we identified 288 students as AR for poor problem-solving outcomes. To derive a parsimonious equation for predicting problem-solving outcomes, we conducted regression analyses on a previous database (L. S. Fuchs, Fuchs, Prentice, et al., 2004) of third-grade students who had received Hot Math SBI. The final prediction equation included pretest performance on the immediate transfer problem-solving measure (see *Measures* section) and pretest performance on the Test of Computational Fluency (L. S. Fuchs et al., 1990). For each cohort, we rank ordered students on the predicted score and selected the lowest 72 students in that year's sample. Within cohort, we as-



signed these students to tutoring conditions, while stratifying by classroom condition. In this way, some AR students received neither classroom nor tutoring Hot Math SBI; some received classroom but not tutoring Hot Math SBI; some received tutoring but not classroom Hot Math SBI; and some received classroom and tutoring Hot Math SBI. Of the 288 AR students, 45 moved prior to posttesting. On demographic and pretest performance variables, the 45 children who moved prior to posttesting were comparable to the 243 pupils who remained in the study, and there were no significant interactions between AR students' tutoring condition and attrition status.

During the first 2 years of the study, AR students were randomly assigned to control (no tutoring) or to one of two versions of tutoring SBI. For Cohort 1, the two versions of SBI tutoring were SBI with and without self-regulated learning strategies; the addition of self-regulated learning strategies appeared promising (the cohort sample was too small to allow us to apply inferential statistics). For Cohort 2, we strengthened the motivation system within self-regulated learning strategies. Students were randomly assigned to receive SBI with strengthened self-regulated learning strategies, with and without drill/practice activities designed to build fluency with low-level skills. Strengthened self-regulated learning strategies appeared promising, but the drill/practice activities did not (again, the cohort sample was too small to allow us to apply inferential statistics). For Cohorts 3 and 4, we built power by randomly assigning AR students to only two tutoring conditions: control versus Hot Math SBI that incorporated strengthened self-regulated learning strategies. In our model, Hot Math SBI tutoring is treated as one condition; however, to gain insight into the effects of the variants, we considered cohort effects. (See Table 1 for a summary of the SBI variants.) Research assistants delivered SBI tutoring, and we refer to these research assistants as *RA tutors*. (When there was more than one variant of SBI tutoring, each RA tutor taught in both conditions.) Information on research assistants is provided later.

For the classroom factor, we had two conditions: classroom control and classroom SBI. For the tutoring factor, we had three conditions: NAR (i.e., not eligible for tutoring), AR control, and AR tutored. See Table 2 for demographics by classroom and tutoring conditions (for a table that provides this information by cohort as well, contact Lynn S. Fuchs). See Table 2 for standard scores on the Wechsler Abbreviated Intelligence Scale IQ, WJ III Applied Problems, and Woodcock Reading Mastery Test-Revised Word Identification, for raw scores on the Test of Computational Fluency (digits correct and problems correct), and for demographic data by classroom and by tutoring conditions (for tables with this information by cohort as well, contact Lynn S. Fuchs). We ran three-way analyses of variance on these performance data, using classroom condition, tutoring condition, and cohort as variables. The only significant effect was for tutoring condition. As expected, on each measure, NAR students scored significantly higher than both AR groups, which scored comparably to each other. We ran chi-square analyses to examine the relation between demographics and classroom condition; none was significant. For tutoring condition, we used one degree of freedom to examine the relation between demographics and risk status. Compared with AR students, NAR students were significantly less likely to receive subsidized lunch, to be African American, and to receive special education (but were comparably likely to be English language

learners). We used the other degree of freedom to examine the relations between demographics and the AR control group versus the AR tutored group; none was significant.

### Classroom Conditions

For a manual with the researcher-designed general problem-solving strategies, classroom SBI, or SBI tutoring teaching scripts and materials, contact Lynn S. Fuchs. Scripts and materials provided a road map for research assistants to implement the instructional methods. The scripts were studied, not read.

*Problem types.* We centered the study on four word-problem types that we chose from the district curriculum to ensure that classroom control students had instruction relevant to the study. The four problem types were *shopping list problems* (e.g., Joe needs supplies for the science project. He needs 2 batteries, 3 wires, and 1 board. Batteries cost \$4 each, wires cost \$2 each, and boards cost \$6 each. How much money does he need to buy supplies?), *half problems* (e.g., Marcy will buy 14 baseball cards. She'll give her brother half the cards. How many cards will Marcy have?), *step-up function or buying bags problems* (e.g., Jose needs 32 party hats for his party. Party Hats come in bags of 4. How many bags of party hats does Jose need?), and *two-step pictograph problems* (e.g., Mary keeps track of the number of chores she does on this chart [a pictograph is shown with a label; each picture stands for 3 chores]. She also took her grandmother to the market 3 times last week. How many chores has Mary done?).

*Classroom control condition.* To guide instruction relevant to the four problem types, classroom control teachers relied primarily on the core math instructional program. For Cohorts 1 and 2, this was Math Advantage (Burton & Maletsky, 1999). For Cohorts 3 and 4, this was Houghton Mifflin Math (Greenes et al., 2005). In both programs, instruction addressed one problem type at a time (as did SBI) and focused on the concepts underlying the problem type. In addition, a prescribed set of problem-solution rules was taught, with explicit steps for arriving at solutions to the problems presented in the narrative. There was no attempt to broaden students' schemas for these problem types to address transfer. However, in comparison to classroom SBI, classroom control group instruction provided more practice in applying problem-solution rules and provided greater emphasis on computational requirements. Classroom control group instruction was explicit and relied on worked examples, guided group practice, independent work with checking, and homework. Compared to Math Advantage, Houghton Mifflin Math had a greater emphasis on problem-solving strategies, including guiding questions to help students understand, plan, solve, and reflect on the content of problems (we note, however, that our analyses did not reveal cohort effects involving the control group).

In addition, the classroom control group (as well as the classroom SBI group) incorporated a 3-week unit on general problem-solving strategies designed and delivered by the researchers (see later discussion) to control for foundational problem-solving strategies that were necessary for but conceptually unrelated to SBI. Previous research (L. S. Fuchs, Fuchs, et al., 2003a) showed that this 3-week unit alone is insufficient to enhance outcomes on the math problem-solving measures used in the present study. We conceptualized the classroom control group as representing conventional classroom problem-solving instruction in order to esti-

Table 2  
Students' Math and Reading Scores and Demographics by Classroom Condition and Tutoring Condition

Variable	Classroom condition												Across tutoring conditions											
	Control: condition						SBI: condition						Across						AR control					
	NAR			AR control			AR control			NAR			Across			NAR			AR control			AR tutor		
	M	SD	(n = 298)	M	SD	(n = 28)	M	SD	(n = 56)	M	SD	(n = 382)	M	SD	(n = 600)	M	SD	(n = 51)	M	SD	(n = 108)	M	SD	(n = 164)
<b>Scores</b>																								
WASI-IQ	99.94	14.19	92.50	11.47	87.59	9.81	97.58	14.18	99.59	13.70	88.25	9.82	87.74	9.72	97.14	13.81	99.70	13.86	89.76	10.56	87.69	9.72		
WJ-AP	105.79	13.05	92.54	12.58	93.75	10.09	103.05	13.62	105.20	13.04	90.47	9.60	91.94	9.57	102.31	13.60	105.39	13.04	91.20	10.72	92.55	9.76		
WRMT-WID	101.87	10.21	95.14	6.45	95.14	6.45	100.42	10.16	102.73	10.00	92.57	5.57	95.60	7.87	101.03	10.05	102.45	10.07	93.48	5.99	95.51	8.21		
Comp Flu dig	24.18	10.02	14.86	7.34	14.25	6.06	22.06	10.18	23.62	9.68	13.55	6.00	14.78	6.45	21.69	9.83	23.80	9.79	14.01	6.49	14.60	6.30		
Comp Flu prob	13.74	5.78	7.14	3.69	6.95	3.57	12.28	6.05	13.47	5.82	6.57	2.93	7.25	3.34	12.13	5.99	13.56	5.81	6.77	13.21	7.15	3.41		
<b>Demographics</b>																								
Sex female	152	51.0	16	57.1	26	46.4	194	50.8	318	53.0	25	49.0	50	46.3	393	51.8	470	52.3	36	45.6	76	48.2		
Subsidized lunch	172	57.7	21	75.0	43	76.8	236	61.2	331	55.2	41	80.4	76	70.4	448	59.0	503	56.0	63	79.7	119	72.6		
<b>Race</b>																								
AA	125	41.9	17	60.7	35	62.5	177	46.3	207	34.5	31	60.8	72	66.7	310	40.8	332	37.0	48	59.5	107	65.2		
EA	123	41.3	6	21.4	16	28.6	145	38.0	278	46.3	15	29.4	30	27.8	323	42.6	401	36.0	21	26.6	46	28.0		
Hisp	33	11.1	3	10.7	2	3.6	38	9.9	71	11.8	5	9.8	6	5.6	82	10.8	104	11.6	8	10.1	8	4.9		
Kurd	4	1.3	0	0.0	0	0.0	4	1.0	13	2.2	0	0.0	0	0.0	13	1.7	17	1.9	0	0.0	0	0.0		
Other	13	4.4	2	7.1	3	5.4	18	4.7	31	5.2	0	0.0	0	0.0	31	4.1	44	4.9	2	2.5	3	1.8		
ELL	13	4.4	1	3.6	0	0.0	14	3.7	19	3.2	2	3.9	2	1.9	23	3.0	32	3.6	3	3.8	2	1.2		
Special education	5	1.7	4	14.3	6	10.7	15	3.9	12	2.0	6	11.8	14	13.0	32	4.2	18	2.0	10	12.7	20	12.3		

Note. Standard scores are given for the WASI-IQ, the WJ-AP, and the WRMT-WID. Raw scores are given for the Comp Flu. SBI = schema-broadening instruction; NAR = not at risk; AR = at risk; WASI-IQ = Wechsler Abbreviated Scale of Intelligence; WJ-AP = Woodcock-Johnson III Tests of Achievement—Applied Problems; WRMT-WID = Woodcock Reading Mastery Tests—Word Identification; Comp. Flu = Test of Computational Fluency; Dig = digits correct; Prob = problems correct; AA = African American; EA = Euro-American; Hisp = Hispanic; ELL = English language learner.



mate the contribution of classroom SBI over conventional instruction.

*General problem-solving strategies instruction (classroom control and classroom SBI groups).* Classroom control and classroom SBI students received a researcher-designed 3-week (two lessons per week) instructional unit on general math problem-solving strategies, which was conceptually unrelated to SBI and was delivered by research assistants. It addressed making sure answers make sense, lining up numbers from text to perform math operations, checking computation, and labeling work with words, monetary signs, and mathematical symbols. These six lessons, each lasting 30–40 min, relied on worked examples with explicit instruction, dyadic practice, independent work with checking, and homework, for a total of 210 min.

*Classroom SBI.* After the classroom-SBI students received the 3-week general math problem-solving unit, they received four researcher-designed 3-week SBI units. Each SBI unit comprised six sessions. Also, two cumulative review sessions were delivered the week after winter break. In each unit, Sessions 1 and 5 lasted about 40 min; the others lasted about 30 min. This totaled 200 min per unit and 856 min across the units (including the two cumulative review sessions). Each 3-week unit addressed one of the four problem types: shopping list, buying bag, half, and pictograph.

Within each unit, the sequence of lessons was as follows. In Sessions 1–4, problem–solution instruction was delivered, with problems that varied only in their cover stories. A poster listing the steps of the solution method was displayed in the classroom. In Session 1, RA teachers addressed the underlying concepts and structural features for the problem type, presented a worked example, and as they referred to the poster, explained how each step of the solution method had been applied in the example. Students responded frequently to questions. After reviewing the concepts and presenting several worked examples in this way, RA teachers shared partially worked examples while students applied the solution steps. Students then completed one to four problems in dyads, where stronger students helped weaker students solve problems and check work with answer keys. Sessions 2–4 were structured similarly, with a greater proportion of time spent on partially worked examples and dyadic practice. Also, at the end of Sessions 2–4, students completed one problem independently. The teacher checked work against an answer key, and students graphed scores.

Sessions 5–6 were designed to broaden schemas, with each problem varying the cover story and one of the four superficial features addressed in SBI. RA teachers first taught the meaning of the word *transfer* and then taught the four superficial features (referred to as *transfer features*), which change a problem without altering its type or solution: A familiar problem type, for which a solution is known, can use unfamiliar vocabulary, can pose an additional question, can incorporate irrelevant information, and/or can combine problem types. A poster, *Transfer: Ways Problems Change*, was displayed. In Session 5, RA teachers explained the poster, illustrating each superficial feature with a worked example. They gradually moved to partially worked examples. Then, students worked in pairs to apply the solution method to problems that varied in their superficial features. In Session 6, RA teachers reviewed the four superficial features, using similar procedures, except that students spent more time working in dyads and then completed a problem independently, scored work against a key, and graphed scores.

*Delivery.* When more than one SBI variant was used for Cohorts 1–3, classroom research assistants had responsibility for classes in both SBI conditions. To ensure comparable mathematics instructional time across conditions, experimental sessions occurred within the confines of the classroom teachers' mathematics instructional block. At the end of the study, classroom teachers reported the number of minutes per week they spent on math, including time on this research project. Mean numbers of math instructional minutes (including researcher-designed general math problem-solving strategy instruction and SBI) were 274.23 ( $SD = 72.96$ ) and 282.50 ( $SD = 80.59$ ) for the classroom control group and for the classroom SBI group, respectively.

*Treatment fidelity.* Prior to the first delivery of each session, RA teachers agreed on the essential information in the script and made a checklist of points. Each session was audiotaped. At the study's end, two research assistants independently listened to tapes while completing the checklist to identify the percentage of points addressed. We sampled tapes so that, within conditions, RA teachers and lesson types were sampled equitably. In class-level control, one or two tapes were sampled per class (for Unit 1); in class-level SBI, six or seven tapes were sampled per class (distributed equally across Units 1–5). Inter-coder agreement, calculated on 20% of the sampled tapes, was 95.3%. The mean percentage of points addressed was 97.20 ( $SD = 3.06$ ) for classroom control and 96.71 ( $SD = 2.37$ ) for classroom SBI,  $F(1, 117) = 0.88$ ,  $p = .351$ .

### *Tutoring Conditions*

NAR and AR control students continued in their classroom conditions without modification, with one third of AR students randomly assigned to the tutoring control condition and two thirds of AR students randomly assigned to tutoring. AR tutored students received SBI tutoring. Tutoring SBI centered on the same four word-problem types.

Tutoring began after the general math problem-solving unit had been implemented, with AR students identified, randomly assigned, and scheduled for tutoring during the general math problem-solving unit. Tutoring addressed the four 3-week SBI units, with three tutoring sessions conducted each week and with two cumulative review sessions delivered the week after winter break. Tutoring groups included 2–4 students, and each session lasted 20–30 min (i.e., 225 min per unit and 940 min across the units, including the two review sessions). Each 3-week unit addressed one of the four problem types: shopping list, buying bag, half, and pictograph.

The content of Hot Math SBI tutoring mirrored the content covered in classroom SBI. However, in tutoring, the most difficult concepts from classroom SBI were targeted, manipulatives were incorporated more frequently, additional scaffolding to support student learning was used, and self-regulated learning strategies with tangible reinforcement (depending on cohort) were incorporated.

During the first five sessions of each unit, problem–solution instruction was delivered, with problems that varied only in their cover stories. A poster listing the steps of the solution method was displayed during the tutoring session. In Session 1, RA tutors used concrete objects to address the underlying concepts and structural features for the problem type. Together with the students, RA tutors worked several examples and, as they referred to the poster



and the concrete objects, explained why and how each step of the solution method had been applied in the examples. Next, students responded frequently to questions as they worked two to four problems together with their tutor. Beginning in Session 2, students then completed one problem independently, which the tutor reviewed and scored.

Sessions 6–9 were designed to broaden schemas, with each problem varying the cover story and one of two transfer features: different question or irrelevant information. RA tutors first taught the meaning of the word *transfer* and then used a poster, *Transfer: To Move*, to teach the two transfer features. Although five bulleted types of transfer features were included on the poster for continuity with the whole-class sessions, the two covered in tutoring sessions were bolded and in larger font for emphasis. In Session 6, RA tutors explained the poster, illustrating both superficial features with working examples. In Sessions 6–9, students still completed a problem independently, although this was never a transfer problem.

For Cohort 1, AR tutored students were randomly assigned to SBI without self-regulated learning strategies (see preceding description) or to SBI with self-regulated learning strategies. Self-regulated learning strategies were designed to help students become more metacognitively, motivationally, and behaviorally active in their own learning (cf. De Corte, Verschaffel, & Eynde, 2000; Zimmerman, 1995). The strategies incorporated the following components. First, RA tutors reminded students to stay on task by working hard, listening carefully, and following directions. RA tutors set timers for three irregular intervals throughout the session (the timing for each session was standardized across tutors); when the timer sounded, each student earned a point if all students in the group were on task. If any student was off task, no student received a point. Second, students received up to three points per session for accurate work, with the task for which accuracy was rewarded varied across sessions (but standardized across RA tutors). Third, students completed one problem, called the Hot Math problem of the day. This problem was scored on a 20-point scale by the RA tutors, according to a specific rubric; students were encouraged to participate in the scoring and to meet or beat their previous day's score. Fourth, following the Hot Math problem of the day, students shaded their Hot Math thermometers with the number of points they had earned over the session. Fifth, students used their Hot Math thermometers to set a goal for next day. Sixth, students were awarded stickers for each point. At the end of each session, students totaled their points and noted their totals on a game board.

In Year 2, self-regulated learning strategies were incorporated into both experimental tutoring-level conditions as part of the regular treatment, with students earning "dollars" instead of stickers and having the opportunity to spend dollars at the Hot Math store each week or to save their dollars. Both tutoring conditions used these strengthened self-regulated learning strategies, but with or without drill-and-practice warm-ups at the beginning of each lesson. These warm-ups were designed to build fluency with a foundational skill necessary for the specific unit (e.g., repeated addition for the stepping-up problem type). Thus, AR tutored students were randomly assigned to SBI with strengthened self-regulated learning strategies or SBI with strengthened self-regulated learning strategies plus warm-ups. All students in Cohorts 3 and 4 who were randomly assigned to tutoring received SBI with strengthened self-regulated learning strategies.

*Delivery.* RA tutors had responsibility for tutoring (for Cohorts 1 and 2, each RA tutored groups in both SBI tutoring conditions). Consistent with an RTI prevention system, tutoring occurred in addition to, not as a substitute for, classroom math instruction and was conducted by someone other than the classroom teacher, in this case the RA tutor. Also, as with an RTI prevention system, students within a given tutoring group were not necessarily from the same classroom.

*Treatment fidelity.* Prior to the first delivery of each session in each condition, RA tutors agreed on the essential information in the script and made a checklist of points. Each tutoring session was audiotaped. At the study's end, four research assistants independently listened to tapes while completing the checklist to identify the percentage of points addressed. We sampled tapes so that, within conditions, RA tutors, groups, and session numbers were sampled equitably. For each of 64 small tutoring groups, 20% of sessions were sampled (seven or eight tapes distributed equally across the four units). Intercooder agreement, calculated on 20% of the sampled tapes, was 96.4%. The mean percentage of points addressed across all units was 98.12 ( $SD = 1.28$ ).

### Research Assistants

Across the 4 years of the study, the typical research assistant was 1 to 2 years beyond undergraduate education, studying for a graduate degree in education, special education, counseling, or education policy. The majority of research assistants worked for the project for 1 year, with six RA teachers remaining for more than 1 year and three RA tutors working for more than 1 year. During each year of the study, two full-time project coordinators, typically with bachelor's- or master's-level degrees outside of education, also served as RA teachers and RA tutors. Each year, five or six RA teachers and another five or six RA tutors were needed. The same research assistants did not conduct classroom SBI and tutoring sessions.

For classroom SBI, a typical RA teacher was responsible for the same four classrooms across 16 weeks of intervention, teaching each classroom two times each week, usually two classes per day, and alternating days for a given classroom. For example, an RA teacher might teach Classrooms 1 and 2 on Monday, Classrooms 3 and 4 on Tuesday, Classrooms 1 and 2 again on Wednesday, and Classrooms 3 and 4 again on Thursday.

Each RA tutor was responsible for one or two control classrooms for general problem-solving strategies during the first 3 weeks and then for three or four tutoring groups during the next 13 weeks. A typical RA tutor taught the same three or four tutoring groups across the 13 weeks, three times per week. The RA tutor's schedule was usually the same on each of the three tutoring days, with students typically receiving tutoring on Tuesday, Wednesday, and Thursday. Tutoring groups were determined on the basis of preferred schedules provided by classroom teachers.

To train research assistants, we conducted the following activities. In an introductory meeting, all research assistants were trained on research ethics and procedures, including professional behaviors in schools. Then research assistants were trained for their specific role within the project (RA teacher or RA tutor) in one full-day session. Research assistants were introduced to the project and its goals, assigned roles (RA teacher or RA tutor), and provided with instruction, demonstrations, and scripted materials



to study. They were paired to practice the treatment. Then, they conducted one lesson for a project coordinator and were judged on a point-by-point system for fidelity to treatment. A research assistant who achieved 95% fidelity was considered reliable. A research assistant who scored lower than 95% fidelity was coached on points he or she missed, asked to practice more, and then rerated at a later time on another lesson. At weekly meetings, project staff met separately with RA teachers and with RA tutors for problem-solving sessions. At the beginning of each unit, a 3-hr training sessions was conducted to orient research assistants and to distribute supporting materials. In addition, all sessions were audiotaped, and the audiotapes were used in part to monitor research assistants for fidelity throughout treatment. Research assistants were provided with written corrective feedback and provided with one-to-one retraining as needed.

### Measures

*Math problem solving.* We used three measures of problem solving. These measures all sampled novel problems (i.e., never seen before or used for instruction). They incorporated the four targeted problem types, but the measures differed from each other in terms of transfer distance in relation to problems used for problem-solution instruction. We refer to the three measures as immediate transfer, near transfer, and far transfer (increasing numbers signify greater transfer distance).

The immediate-transfer measure incorporated novel problems in the same format as the problems used for problem-solution instruction. The near-transfer measure incorporated novel problems that varied from the problems used for problem-solution instruction in terms of one or more of the transfer features addressed in SBI: unfamiliar vocabulary, different question, irrelevant information, or combination of problem types. The far-transfer measure, which was designed to mirror real-life problems, varied from the problems used for instruction in multiple ways. The far-transfer measure was formatted to look like a commercial, standardized test; it presented a multiparagraph narrative with four questions; some of the information needed to answer the question was removed from the multiparagraph narrative and placed in figures or question stems; it contained multiple pieces of numeric and narrative irrelevant information; it provided opportunities for students to formulate decisions; it combined all four problem types; and it varied all four SBI transfer features.

None of the cover stories had been used for instruction. Each measure had two alternate forms; problems on both forms required the same operations and presented text with the same number/length of words. Immediate-transfer alternate forms incorporated the same numbers; near-transfer alternate forms incorporated the same numbers; and far-transfer alternate forms used similar numbers. In half the classes in each condition, we used Form A at pretest and Form B at posttest; in the other half, forms were reversed.

The immediate-transfer measure comprised 10 word problems, equally representing the problem-solving units. Across the 10 problems, the maximum score was 44. For this sample, Cronbach's alpha was .84 at pretest and .95 at posttest; concurrent validity with WJ III Applied Problems (Woodcock et al., 2001) was .56 at pretest and .42 at posttest. Interscorer agreement, computed on

20% of protocols by two independent scorers who were unaware of the purpose of the study, was .983 at pretest and .960 at posttest.

The near-transfer measure comprised nine problems: a shopping list problem with a novel format (information shown in bulleted format, with a selection rather than an open-ended response format); a shopping list problem with a novel question (asking for money left at the end); a buying bags problem with a different keyword (*packages* instead of *bags*); a buying bags problem with a novel question (comparing prices of two packaging options); a half problem with unfamiliar vocabulary (*share equally* instead of *half*); a pictograph problem with a novel question (asking for money left at the end); a pictograph problem with a novel question (comparing quantities at the end); a problem with irrelevant information that combined a buying bags problem with a pictograph problem and combined novel vocabulary with a novel question; and a problem with irrelevant information that combined a shopping list problem with a buying bags problem and combined a novel format with a novel question. For this sample, across items, the maximum score was 79. Cronbach's alpha was .87 at pretest and .96 at posttest; concurrent validity with WJ III Applied Problems (Woodcock et al., 2001) was .52 at pretest and .40 at posttest. Interscorer agreement, computed on 20% of protocols by two independent scorers who were unaware of the purpose of the study, was .985 at pretest and .973 at posttest.

The far-transfer measure simultaneously assessed transfer of all four problem types and the four transfer features addressed in SBI. Also, to decrease association between the task and classroom or tutoring SBI, the far-transfer measure was formatted to look like a commercial test (printed with a formal cover, on green paper, with photographs and graphics interspersed throughout the test booklet). Two assessments were constructed as alternate forms: Although the context of the problem situations differed, the structure of the problem situation and the questions were identical, and the problem solutions and reading demands were equivalent.

Performance was scored according to a rubric with four dimensions: conceptual underpinnings, computational applications, problem-solving strategies, and communicative value. The original rubric (Kansas Board of Education, 1991) scored responses on a 6-point scale. To enhance reliability, we awarded points on a finer basis (e.g., the problem-solving strategies score included points for finding relevant information, accumulating to a total, showing all computation, working the answer in distinct multiple parts, labeling at least half of the multiple parts, and labeling work with monetary and operation signs). Across the four questions and four scoring dimensions, the maximum score is 72. For this sample, Cronbach's alpha was .91 at pretest and .94 at posttest; concurrent validity with WJ III Applied Problems (Woodcock et al., 2001) was .47 at pretest and .61 at posttest. Interscorer agreement, computed on 20% of protocols by two independent scorers who were unaware of the purpose of the study, was .987 at pretest and .949 at posttest. Given the deleterious effects of student unfamiliarity with performance assessments (L. S. Fuchs et al., 2000), research assistants delivered a 45-min lesson on test taking strategies before pre- and posttesting in all conditions. (The mean score across immediate, near, and far transfer correlated .62 with WJ III Applied Problems, Woodcock et al., 2001, at pretest and .57 at posttest.)

*Math applications.* The 60-item WJ III Applied Problems (Woodcock et al., 2001) measures skill in analyzing and solving



practical math problems. The tester orally presents items involving counting, telling time or temperature, and problem solving. Testing is discontinued after six consecutive errors. The score is the number of correct items. As reported by McGrew and Woodcock (2001), the 1-year test-retest reliability is .85; the ratio of true score variance to observed variance is .88–.91. Coefficient alpha on this sample was .85.

### *Data Collection*

Trained research assistants administered the three problem-solving measures in whole-class arrangement. For the immediate-transfer and near-transfer measures, research assistants read aloud each item and provided students with time to complete their work before progressing to the next item. For the far-transfer measure, research assistants read the entire assessment and reread portions to individuals, at their request, as they worked. Trained research assistants administered the WJ III Applied Problems individually, outside the classroom in a quiet location in the school. All pre-testing occurred during the 3 weeks before treatment; posttesting occurred during the 3 weeks following treatment. To avoid having familiar research assistants prompting awareness that SBI might apply, students were unfamiliar with their testers. The RA teachers and RA tutors used scripted sets of directions to administer the three whole-class problem-solving measures. In many cases, the RA teachers and RA tutors could identify study conditions, but given the whole-class, scripted administration, it seems unlikely that testers could have influenced performance. An entirely different pool of research assistants administered individual WJ III Applied Problems, so these testers had no way of knowing the study conditions in which students had participated.

### *Data Analysis*

We converted scores on the three problem-solving measures to percentage correct so that performance on the three measures could be compared. To examine how much of the total variance in improvement on the three problem-solving measures was explained by the clustering of children in classrooms and in tutoring groups, we estimated variance components with SAS PROC MIXED (Littell, Milliken, Stroup, Wolfinger, & Schabenger, 2006). The resulting intraclass correlations showed that the effect for classroom clustering explained 16.10% of the variance ( $p < .001$ ) and that the effect for tutoring-group clustering explained 4.10% of the variance ( $p = .006$ ). We therefore incorporated each as a random effect into our model, which also included four fixed effects: one within-subjects factor (problem-solving measure) and three between-subjects factors (classroom condition, tutoring condition, and cohort). To assess pretreatment comparability, we fit a full model that included all main effects, two-way, three-way, and four-way interactions as well as estimating the impact of classroom as a random effects factor.

To index learning as a function of study condition, we used improvement on the three problem-solving measures. (Fitting a model with improvement scores produces the same effects as would considering the interaction between test occasion [pre- vs. posttest] and study conditions. We opted for improvement scores because their interpretation is more straightforward.) In this full model, the variance component for tutoring group decreased to

zero (indicating that all of the variance associated with tutoring group clusters was explained in the model). Therefore, we fixed the random effects of tutoring group to zero and also eliminated from the final model all higher order interactions that were not statistically significant. To follow up significant effects, we Bonferroni corrected  $p$  values by the number of follow-up tests we ran for that significant effect. We subtracted the difference between improvement means and then divided by the pooled standard deviation of the improvement/square root of  $2(1 - r_{xy})$  to compute effect sizes (ESs; Glass, McGaw, & Smith, 1981). To examine the effects of the classroom and tutoring tiers of SBI prevention on the prevalence of math difficulty, we designated difficulty as performing below the 16th percentile on WJ III Applied Problems at the end of intervention and applied chi-square analyses.

## Results

See Table 3 for means and standard deviations by classroom condition and by tutoring condition (for means and standard deviations by classroom condition, tutoring condition, and cohort, contact Lynn S. Fuchs). See Table 3 for rates of difficulty status by classroom and tutoring conditions and for ESs.

### *Pretreatment Comparability on Math Problem Solving*

See Table 3 for means and standard deviations. There were no significant effects for classroom condition,  $F(1, 274) = 0.06$ ,  $p = .811$ , or for cohort,  $F(3, 3274) = 0.14$ ,  $p = .938$ . As expected, there was a significant effect for tutoring condition,  $F(2, 3274) = 236.75$ ,  $p < .001$ , with NAR students performing higher than each of the two AR tutoring groups (both  $p < .001$ ), which performed comparably to each other ( $p = .587$ ). Also, there was a significant effect for measure,  $F(2, 3274) = 27.43$ ,  $p < .001$ . Across classroom conditions, across tutoring conditions, and across cohorts, students performed higher on immediate transfer than on near transfer, performed higher on immediate transfer than on far transfer ( $p < .001$ ), but performed higher on far transfer than on near transfer ( $p < .001$ ). Students scored higher on far than on near transfer at pretreatment because far transfer incorporated a greater variety of problem types, some of which were simpler than the problem types on near transfer. Neither of these main effects threatens the internal validity of the study. (Moreover, as shown later, improvement on far transfer was harder to effect, supporting its far-transfer designation.) The two-way interactions were as follows: for the Classroom Condition  $\times$  Tutoring Condition interaction,  $F(2, 3274) = 0.22$ ,  $p < .799$ ; for the Classroom Condition  $\times$  Measure interaction,  $F(2, 3274) = 0.21$ ,  $p < .814$ ; for the Tutoring Condition  $\times$  Measure interaction,  $F(4, 3274) = 28.75$ ,  $p < .001$ ; for the Classroom Condition  $\times$  Cohort interaction,  $F(3, 3274) = 0.52$ ,  $p < .666$ ; for the Tutoring Condition  $\times$  Cohort interaction,  $F(6, 3274) = 0.68$ ,  $p < .663$ ; and for the Measure  $\times$  Cohort interaction,  $F(6, 3274) = 2.76$ ,  $p = .011$ . For Cohorts 1, 2, and 4, students performed higher on immediate transfer than on near transfer (all  $p < .001$ ), performed higher on immediate transfer than on far transfer (all  $p < .001$ ), but performed higher on far transfer than on near transfer (all  $ps < .001$ ; except for Cohort 1,  $p = .005$ ). By contrast, the third cohort scored comparably on near and far transfer ( $p = .022$ ; with immediate-transfer scores higher than both other measures, as was the case for the other three



Table 3  
Performance on Problem-Solving Measures by Classroom Condition and Tutoring Condition (Across Cohorts)

Measure	Class-level condition											
	Control: Tutoring-level condition						SBI: Tutoring-level condition					
	NAR		AR control		AR tutor		NAR		AR control		AR tutor	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Immediate transfer	21.94	13.62	4.25	3.20	5.48	4.93	18.23	14.06	4.34	4.48	4.05	18.64
Pretreatment	41.55	21.55	22.30	17.32	45.41	21.24	40.71	21.84	46.18	72.82	23.64	77.31
Posttreatment	19.61	18.75	18.11	18.09	39.92	21.87	22.48	20.45	41.83	68.34	23.65	58.67
Improvement												
Near transfer	9.14	7.91	3.12	3.22	4.27	3.74	7.99	7.51	2.01	2.70	1.63	7.50
Pretreatment	21.22	14.18	9.27	8.56	21.77	14.66	20.43	14.24	21.97	37.46	14.46	45.60
Posttreatment	12.08	12.11	6.15	7.66	17.50	13.45	12.44	12.03	19.96	35.76	13.16	38.10
Improvement												
Far transfer	14.05	14.45	6.42	6.86	6.26	6.11	12.35	13.49	5.34	6.30	6.38	12.73
Pretreatment	30.84	19.95	15.75	11.17	20.42	12.86	28.21	19.20	15.81	20.37	14.79	35.30
Posttreatment	16.79	17.56	9.33	12.33	14.16	13.48	15.86	16.78	10.47	11.22	13.50	22.57
Improvement												
Across	15.04	9.97	4.60	3.03	5.34	3.38	12.86	9.84	3.90	4.49	3.19	12.95
Pretreatment	31.21	16.04	15.97	10.42	29.20	14.28	29.78	15.96	27.98	43.55	14.23	52.74
Posttreatment	16.16	12.03	11.20	9.79	23.86	13.50	16.93	12.49	24.09	39.06	13.10	39.78
Improvement												
Across class-level conditions												
NAR												
(n = 298)												
AR control												
(n = 28)												
AR tutor												
(n = 56)												
Across												
(n = 382)												
NAR												
(n = 600)												
AR control												
(n = 51)												
AR tutor												
(n = 108)												
Across												
(n = 759)												
NAR												
(n = 898)												
AR control												
(n = 79)												
AR tutor												
(n = 164)												

Note. SBI = schema-broadening instruction; NAR = not at risk; AR = at risk.

cohorts, both  $ps < .001$ ). Because neither classroom nor tutoring conditions were involved in this interaction, it does not threaten the validity of the study. None of the three- or four-way interactions was significant.

### *Math Problem-Solving Learning as a Function of Prevention Conditions*

See Table 3 for means and standard deviations; see Table 4 for ESs. We first consider the primary question addressed in the present study: Is tutoring more effective with validated classroom instruction? Then, we consider other effects involving treatment. Finally, we list results for the remaining effects, which are of less interest.

*Is tutoring more effective with validated classroom instruction?* The interaction between classroom condition and tutoring condition was significant,  $F(2, 3320) = 24.33, p < .001$ . We pursued this interaction in three ways. First, exclusively with AR students, we tested the difference between tutoring with and without validated classroom instruction directly; this was significant ( $p < .001$ ;  $ES = 1.34$ ). Second, again focusing exclusively on AR learners, we contrasted the difference between tutored and control students who received conventional classroom instruction ( $ES = 1.18$ ) versus the difference between tutored and control students who received validated SBI classroom instruction ( $ES = 1.13$ ). This was not significant ( $p = .308$ ), suggesting that the effects of two tiers of prevention are additive, not synergistic. Finally, we considered the pattern of improvement between NAR versus AR students as a function of tutoring condition within each classroom treatment. With classroom control, improvement of NAR students and AR control students was comparable ( $p = .011$ ); NAR students improved less than AR tutored students ( $p < .001$ ); and AR tutored students improved more than AR controls ( $p < .001$ ). By contrast, when students received classroom SBI, NAR students outgrew AR control students ( $p < .001$ ); AR tutored students continued to outgrow AR controls ( $p < .001$ ); and the improvement of NAR and AR tutored students was comparable ( $p = .108$ ).

*Interactions with problem-solving measure.* The main effect for classroom condition was significant,  $F(1, 3320) = 160.32, p < .001$ . Classroom condition did, however, interact significantly with measure,  $F(2, 3320) = 176.37, p < .001$ . The effect of classroom condition was significant on immediate transfer ( $p < .001$ ) and on near transfer ( $p < .001$ ), but not on far transfer ( $p = .495$ ).

The main effect of tutoring condition was also significant,  $F(2, 3320) = 51.67, p < .001$ . There was, however, a significant interaction between tutoring condition and measure,  $F(4, 3320) = 27.61, p < .001$ . On immediate transfer, NAR students outgrew AR control students ( $p < .001$ ). Of importance, however, AR tutored students outgrew AR control students ( $p < .001$ ), and most impressively, AR tutored students also outgrew NAR students ( $p < .001$ ). On near transfer, although NAR and the AR tutored students again outgrew AR control students (both  $ps < .001$ ), the improvement of the NAR and the AR tutored students was comparable on this measure ( $p = .813$ ). A different pattern emerged on the far-transfer measure. Although the contrast between AR tutored students and AR control students only approached significance ( $p < .043$ ), the NAR students outgrew both AR groups of learners (both  $ps < .001$ ).

*Interaction between tutoring and cohort.* Tutoring also interacted significantly with cohort,  $F(6, 3320) = 3.45, p = .002$ . In the follow-up to this interaction, we focused exclusively on AR students by looking at the effect of receiving versus not receiving tutoring for Cohorts 2, 3, and 4 and comparing each to the effect for Cohort 1. This is interesting because with tutoring SBI, none of the students in Cohort 1 had the strengthened version of self-regulated learning strategies, and although all of Cohort 2 had the strengthened version of self-regulated learning strategies, half of this cohort had the warm-up drill and practice activities. By contrast, all of the SBI tutored students in Cohorts 3 and 4 received the strengthened version of self-regulated learning strategies without warm-up activities. In line with this instructional design feature, the contrast between tutored Cohorts 1 and 2 was not significant ( $p = .194$ ; ESs comparing AR tutored against AR control were

Table 4  
*Effect Sizes on Improvement Scores by Classroom Condition and Tutoring Condition (Across, Cohorts)*

Contrast	Measure			
	Immediate transfer	Near transfer	Far transfer	Across
Within-classroom control				
NAR vs. AR tutored	-1.16	-0.42	0.15	-0.50
NAR vs. AR control	0.09	0.51	0.43	0.37
AR tutored vs. AR control	1.47	1.03	0.49	1.18
Within-classroom SBI				
NAR vs. AR tutored	-0.46	0.36	0.86	0.14
NAR vs. AR control	0.80	1.73	1.23	1.17
AR tutored vs. AR control	1.52	0.99	0.23	1.13
Across-classroom conditions				
NAR vs. AR tutored	-0.54	0.11	0.44	-0.05
NAR vs. AR control	0.53	0.86	0.64	0.73
AR tutored vs. AR control	1.34	1.09	0.37	1.17
Across-tutoring conditions: SBI vs. control	1.90	1.73	0.36	1.51

*Note.* A positive value indicates that the first condition named was higher. A negative value indicates that the second condition named was higher. NAR = not at risk; AR = at risk; SBI = schema-broadening instruction.



0.60 for Cohort 1 vs. 0.97 for Cohort 2), but the contrast between Cohorts 1 and 3 was significant ( $p < .001$ ; ESs comparing AR tutored against AR control were 0.60 for Cohort 1 vs. 1.86 for Cohort 3), the contrast and between Cohorts 1 and 4 was also significant ( $p = .002$ ; ESs comparing AR tutored against AR control were 0.60 for Cohort 1 vs. 1.57 for Cohort 4).

*Remaining effects.* The measure effect was significant,  $F(2, 3320) = 219.97, p < .001$ . As expected, across classroom conditions, tutoring conditions, and cohorts, improvement on immediate transfer exceeded improvement on near transfer ( $p < .001$ ); improvement on immediate transfer exceeded improvement on far transfer ( $p < .001$ ); and improvement on near transfer exceeded improvement on far transfer ( $p < .001$ ). The cohort effect was not significant,  $F(1, 3320) = 0.43, p < .733$  (indicating that across classroom conditions, tutoring conditions, and measures, improvement of the four cohorts was comparable).

Math Difficulty as a Function of Prevention Conditions

See Table 5 for math difficulty by classroom and tutoring conditions. Chi-square analyses indicated no relation between math difficulty status and classroom condition,  $\chi^2(1, N = 1141) = 0.18, p = .672$ . To examine the relation between math difficulty status and tutoring condition (where we had three levels: NAR, AR control, and AR tutored), we allocated one degree of freedom to the contrast between NAR versus AR (control and tutored), which revealed a significant relation:  $\chi^2(1, N = 1141) = 104.27, p < .001$ . We allocated the other degree of freedom to the contrast between the two AR groups, which also revealed a significant relation:  $\chi^2(1, N = 243) = 7.08, p = .008$ .

We then estimated overall prevalence of math difficulty as a function of whether AR students received tutoring. To estimate overall prevalence without tutoring, we did the following calculation: (the number of NAR students identified as having math difficulty plus the number of AR control students identified as having math difficulty plus [the proportion of AR control students identified as having math difficulty times the proportion of AR tutored students]) divided by the number of students in the 89 classrooms who completed the study ( $n = 1,141$ ). We added the proportion of AR control students identified as having math difficulty times the proportion of AR tutored students, because we

wanted to include the base rate of expected math difficulty in the tutored group if they had not been tutored. In a similar way, to estimate overall prevalence with tutoring, we did the following calculation: (the number of NAR students identified as having math difficulty plus the number of AR tutored students identified as having math difficulty plus [the proportion of AR tutored students identified as having math difficulty times the proportion of AR control students]) divided by the number of students in the 89 classrooms who completed the study ( $n = 1141$ ). Here, we added the proportion of AR tutored students identified as having math difficulty times the proportion of AR control students, because we wanted to include the rate of math difficulty expected in the control group if they had been tutored. Without tutoring, we estimate a math difficulty prevalence rate of 6.8%; with tutoring, we estimate a rate of 3.9%.

Discussion

The primary purpose of this large-scale randomized field trial was to assess the effects of third-grade small-group tutoring with and without validated classroom instruction on AR students' math problem solving and their learning relative to NAR peers. The issue of how tutoring interacts with classroom instruction among students at risk for poor learning outcomes is important for designing efficient and effective RTI prevention systems. If tutoring is differentially efficacious when combined with validated classroom instruction, then both tiers are critical, and classroom instruction needs to be designed deliberately with AR students in mind, even when they receive tutoring. By contrast, if tutoring promotes comparable outcomes regardless of the classroom instructional context, then findings would set the stage for research on whether tutoring might occur as a replacement for, rather than as a supplement to, classroom instruction. This would make RTI prevention systems logistically easier and more efficient, thus facilitating scheduling and permitting resources to be infused at the tutoring tier. In addition, when considering tutoring efficacy as a function of the classroom context, it is important not only to compare the response of AR students who do and do not receive tutoring, but also to contrast the learning of AR students against their NAR peers. This has not been addressed in prior work.

Therefore, in the present study, we randomly assigned classrooms to validated or conventional instruction, designated incoming risk status (NAR vs. AR), and within classroom conditions, randomly assigned AR students to receive or not receive tutoring. This allowed us to compare the learning of AR students (with one, two, or no tiers of SBI) and NAR students (with and without validated classroom prevention). We considered this an efficacy study, in which the key question was whether interventions had their intended effects when they were actually implemented. Toward that end, we relied on research assistants to implement SBI classroom instruction and SBI tutoring. We note, however, that within most RTI models, someone other than the classroom teacher implements Tier 2 tutoring; therefore, the use of research assistants to conduct tutoring does not compromise external validity.

Results revealed an interaction between the two tiers of prevention. Of importance, tutoring was significantly and substantially more effective when it occurred in combination with validated classroom instruction than when the tutoring occurred with con-

Table 5  
Difficulty Status by Classroom Condition and Tutoring Condition (Across Cohorts)

Tutoring condition for groups with MD <sup>a</sup>	Classroom condition					
	Control		SBI		Across	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
NAR	2	0.6	11	1.8	13	1.4
AR control	8	28.5	13	25.5	21	26.6
AR tutoring	7	12.5	14	12.9	21	12.8
Across	17	4.5	38	5.0	55	4.8

Note. SBI = schema-broadening instruction; MD = math difficulty; NAR = not at risk; AR = at risk.  
<sup>a</sup>Final Woodcock-Johnson Applied Problems score was below the 16th percentile.



ventional classroom instruction ( $ES = 1.34$ ). This suggests that two tiers are better than one tier of prevention and indicates the importance of providing AR students with validated instruction in the classroom and then supplementing that instruction with high-quality tutoring. In the present study, the two tiers of instruction were closely aligned, both addressing the same types of word problems at the same time and both relying on the same theoretical (SBI) and operational (Hot Math) approach to instruction. It is possible that when the two tiers of instruction are less well aligned, as is often the case, results would differ. Moreover, it is possible that alignment of Tier 1 and Tier 2 instructional content may differ as a function of academic domain. Consequently, we emphasize that future studies should assess the added value of validated classroom instruction that is more and less aligned with tutoring for different academic content. At the same time, even with closely aligned instruction at the two tiers of the prevention system, our results showed that the advantage for AR tutored students over AR control students was comparable with ( $ESs = 1.18$ ) and without ( $ES = 1.13$ ) validated classroom instruction. Thus, the effects of two tiers of prevention appear to be additive rather than synergistic.

It is also interesting to consider the effects of one versus two tiers of validated prevention by comparing the performance of AR students against NAR peers (at least in the area within which intervention occurred, i.e., in math problem solving). When contrasting AR tutored students' learning with that of NAR peers, two tiers of prevention again appear to have made a difference, but in an unanticipated way. With only one tier of prevention (i.e., tutoring for AR students combined with conventional classroom instruction for AR and NAR students), AR tutored students improved more than NAR students, by an impressive half a standard deviation. Therefore, when AR and NAR students received conventional classroom math problem-solving instruction, with AR students receiving tutoring, the achievement gap between AR and NAR students narrowed substantially (at preintervention,  $ES = 1.09$ , with the mean difference between AR and NAR students representing 2.80 standard errors of measurement; at postintervention,  $ES = 0.13$ , with the mean difference between AR and NAR students less than 1 standard error of measurement). This unexpected finding shows the power of SBI tutoring to effect strong learning among AR students, who initially performed substantially below NAR classmates. It also suggests that classroom teachers require guidance about how to promote math problem-solving learning. On the other hand, with two tiers of prevention (i.e., validated classroom instruction and tutoring for AR students combined with validated classroom instruction for NAR students), AR tutored and NAR students achieved comparably ( $ES = 0.14$ ). When AR and NAR students received validated classroom math problem-solving instruction, with AR students also receiving a second tier of tutoring, the achievement gap between AR and NAR students remained sizeable and comparable (at preintervention,  $ES = 1.29$ , with the mean difference between AR and NAR students representing 3.11 standard errors of measurement; at postintervention,  $ES = 0.72$ , with the mean difference between AR and NAR students representing 2.71 standard errors of measurement).

In terms of the contrast between AR control and NAR students, NAR students consistently improved more than AR controls, with an  $ES$  of 0.37 without classroom SBI and an  $ES$  of 1.17 with

classroom SBI. This was expected. When AR and NAR students received the same single tier of validated classroom math problem-solving instruction, without AR students receiving a second tier of preventive tutoring, the achievement gap between AR and NAR students grew (at preintervention,  $ES = 1.30$ , with the mean difference between AR and NAR students representing 3.28 standard errors of measurement; at postintervention,  $ES = 1.55$ , with the mean difference between AR and NAR students representing 5.95 standard errors of measurement; hence, the gap between AR and NAR students at pre- versus postintervention increased by 2.69 standard errors of measurement). The differential effect of validated classroom practices, favoring NAR over AR students, has been documented in prior work (e.g., L. S. Fuchs et al., 1997).

Together, findings indicate that intensive instruction, in the form of preventative tutoring, is essential for AR students. Without it, the gap between AR and NAR students continues to widen, even when NAR students suffer the disappointing effects of conventional classroom instruction in math problem solving. Accordingly, results highlight the importance of validated problem-solving instruction in the Tier 1 classroom and suggest that tutoring occur as a supplement to, not a replacement for, classroom instruction. This is the case for AR students: When they receive tutoring combined with validated classroom instruction, their learning exceeds that of students who receive tutoring without validated classroom instruction—by a practically important  $ES$  of 1.34 standard deviations. At the same time, the importance of validated classroom instruction is also clear for NAR students, whose problem-solving learning is superior with validated classroom SBI. Readers should take note, however, that when AR students receive two tiers of validated instruction while their NAR classmates receive validated classroom instruction, the AR students' performance gap is unlikely to narrow. We also caution that these results cannot be generalized beyond third grade, beyond math problem solving, and beyond instruction that is closely aligned across tiers. Additional research exploring these issues at other grades and in other academic areas and research testing the effects with and without instructional alignment are warranted.

This need for additional studies at other grades and in other areas is illustrated by comparing our results to those of Gilbert et al. (2007). Working with reading in first grade, Gilbert et al. showed that validated classroom instruction did not contribute to tutoring efficacy. At least two differences between our study and Gilbert et al.'s study help explain the inconsistency of findings. First, because Gilbert et al. tutored students at instructional levels that were substantially below classmates' levels, classroom teachers and tutors did not work on identical skills at the same time (as was the case in the present study). Second, Gilbert et al.'s AR sample had already proven unresponsive to the fall semester's validated classroom instruction; therefore, it is not surprising that second-semester Tier 2 tutoring was not differentially effective when combined with Tier 1 validated instruction. By contrast, we used a more typical RTI method for identifying students for risk and for tutoring: initial performance on predictors of outcome.

In addition to studying the contribution of validated classroom instruction to the efficacy of tutoring in terms of math problem-solving outcomes, we also examined the prevalence of mathematics difficulty with and without validated classroom instruction and with and without tutoring. We operationalized math difficulty with a widely used tool for identifying math disability in the schools.



Also, WJ III Applied Problems (Woodcock et al., 2001) is broader than our math problem-solving outcome measures, assessing skill in analyzing and solving practical math problems that involve counting and telling time or temperature, as well as problem solving. In addition, we selected a relatively stringent cutpoint (the 15th percentile) for designating difficulty, and we note that students were not identified for participation and/or designated as AR with this measure. Even so, the prevalence of math difficulty was clearly associated with tutoring. Although AR students were, as expected, disproportionately represented in the math difficulty category compared with NAR students, AR tutored students were significantly and dramatically less likely than AR controls to be designated as having math difficulty at the end of the study. That is, tutoring cut the proportion of students with math difficulty in half: from 26.6% of AR control students (or 6.8% of the general population) to 12.8% of AR tutored students (or 3.9% of the general population). At the same time, math difficulty status was not associated with classroom condition, even when looking exclusively at AR students who received validated classroom SBI versus conventional classroom instruction. Thus, it appears that tutoring, not validated classroom instruction, is the active agent that reduces math difficulty status for AR students. This suggests that tutoring is the tier that moves student ranking around the cutpoints used to designate disability and corroborates the strong need for tutoring among the population of learners at risk for poor math problem-solving learning.

Four additional findings, although not the primary purpose of the present study, also merit discussion. First, results strengthen previous work (L. S. Fuchs, Fuchs, et al., 2003b; L. S. Fuchs, Fuchs, Prentice, et al., 2004) showing that Hot Math SBI enhances mathematical problem solving. Across the three problem-solving measures and across tutoring conditions, the ES favoring classroom SBI over conventional teachers' mathematics problem-solving instruction was 1.51 standard deviations. Thus, SBI was superior to teachers' conventional instruction, even when the problem types addressed within SBI were selected to correspond with the schools' (i.e., the control group's) curriculum. Given that the researchers designed SBI with considerable effort over years, it may seem unsurprising that SBI was more effective than conventional classroom instruction. Therefore, it is important to note that L. S. Fuchs, Fuchs, et al. (2003b) demonstrated that SBI makes a sizeable contribution beyond researcher-designed problem-solution instruction, which does not teach students to broaden schemas for transfer to novel problems. We therefore conclude that effects are due to schema-broadening instruction and not simply to researcher-designed problem-solution instruction.

Second, findings demonstrate the efficacy of SBI for tutoring AR students on problem types that extend beyond total, difference, and change relationships addressed in prior work (L. S. Fuchs et al., 2008; Jitendra et al., 1998, 2002; Jitendra & Hoff, 1996; Xin et al., 2005). By contrast, the problem types addressed in the present study involve multiple quantities of more than one kind of item, halving quantities, step-up functions, and pictograph problems, with emphasis on multistep problem solving, combining problem types, and ignoring irrelevant information. Across the three problem-solving measures and across classroom SBI and control conditions, the ES favoring SBI tutoring over control among AR learners exceeded 1 standard deviation ( $ES = 1.17$ ), even though

two thirds of untutored (control) AR students received SBI at the classroom level. Few validated tutoring protocols exist at the second tier of RTI multitier prevention systems. Results provide an option for schools to use within their RTI systems for math problem solving on problem types appropriate for third grade.

Third, the interaction between tutoring condition and cohort suggests that for AR learners, systematic motivation may be important for promoting learning within small-group tutoring. For Cohort 1, although tutoring incorporated self-regulated learning strategies, the motivation system within those self-regulated learning strategies was weak. By contrast, for Cohorts 3 and 4, tutoring incorporated a strengthened motivation system within self-regulated learning strategies, which relied on tangible reinforcers for listening carefully, following directions, paying attention, working hard, and producing accurate work. Accordingly, the differences in outcome for Cohort 1 versus Cohort 3 and for Cohort 1 versus Cohort 4 were statistically significant, with ESs of 1.86 and 1.57, respectively. (This strengthened motivation system was first incorporated with the Cohort 2 tutored sample; however, half of the Cohort 2 tutored sample also received a warm-up drill and practice, which proved problematic. The difference in learning outcome for Cohort 1 vs. Cohort 2 was not statistically significant, even though the ES was a sizeable 0.60.) These findings suggest the need for systematic motivation systems to address task persistence deficits among AR learners (Deci & Chandler, 1986; Deci & Ryan, 1985; Garber & Seligman, 1980). This may be particularly relevant when solving mathematical word problems, which requires meta-cognition and persistence in the face of challenge (De Corte et al., 2000). Causal-comparative studies (e.g., Lester & Garofalo, 1982; Schoenfeld, 1992; Silver, Branca, & Adams, 1980) illustrate how self-regulation differs between weak and skilled problem solvers: Experts spend more time analyzing problems before initiating solutions, reflect more frequently on their problem solving, and alter their approach more flexibly. Motivation to think through challenging tasks and to persist with the task even when initial efforts are not successful may be critical, and future research should address this question more directly.

Fourth, results illustrate the difficulty of enhancing students' real-life problem solving, as revealed with a significant interaction between classroom condition and problem-solving measure. Learning was significantly stronger for classroom SBI compared to classroom control on immediate- and near-transfer measures, both of which included entirely novel math problems, with near-transfer problems extending well beyond what had been directly taught within SBI. The ESs for the immediate- and near-transfer measures, respectively, were 1.90 and 1.73. On far transfer, however, the effect was not statistically significant. The far-transfer measure, designed to mirror real-life problems, varied from the problems used for instruction in multiple ways. This measure presented a multi-paragraph narrative with four questions; some of the information needed to answer the question was removed from the multi-paragraph narrative and placed in figures or question stems; it contained multiple pieces of irrelevant numeric and narrative information; it provided opportunities for students to formulate decisions; it combined all four problem types; and it varied all four SBI transfer features. Given the transfer distance involved in the far-transfer measure, the ES of 0.36 is notable, especially because effects were demonstrated with relatively



young students for whom schema-induction competence (e.g., Chen, 1999) and problem-solving competence (Chen, 1999; Durnin et al., 1997; Foxman et al., 1991, cited in Boaler, 1993; Larkin, 1989) have proven hard to effect. Present findings do, however, suggest the need for additional research to examine how to strengthen real-life problem solving.

In a similar way, an interaction between problem-solving measure and tutoring condition was also documented. On immediate transfer, AR tutored students improved more than NAR students ( $ES = 0.54$ ), and both groups grew more than the AR control group ( $ES = 1.34$  for AR tutored and  $0.53$  for NAR). On near transfer, the improvement of NAR and AR tutored students was comparable ( $ES = 0.11$ ), but again both groups grew better than AR controls ( $ES = 0.86$  for NAR and  $1.09$  for AR tutored). Only on far transfer did the improvement of NAR students exceed that of AR tutored students ( $ES = 0.44$ ), and the improvement of AR tutored versus AR control students only approached statistical significance ( $p = .043$ ;  $ES = 0.37$ ). Again, given the transfer distance involved in the far-transfer measure and the sizeable deficits the AR students brought to the tutoring table, the  $ES$  of  $0.37$  favoring AR tutored over AR control students appears promising, but additional work exploring methods for promoting real-life problem solving is required.

In summary, relying on a randomized field trial while accounting for the nested structure of the data, the present study provides evidence for the efficacy of SBI for classroom use and for tutoring use in promoting superior math problem solving with AR and NAR students. Moreover, in terms of RTI prevention systems, results indicate that two tiers of validated instruction are better than one tier but that the effects of two tiers are additive, not synergistic. Findings also illustrate the disappointing efficacy of conventional math problem-solving instruction and the need for improved methods to teach math problem solving in general education classrooms. SBI appears to represent one potentially strong method for achieving that goal. At the same time, as shown previously (e.g., Bransford & Schwartz, 1999; A. L. Brown et al., 1992; Durnin et al., 1997; Foxman et al., 1991, cited in Boaler, 1993; Larkin, 1989; Mayer et al., 1999), findings once again corroborate the difficulty of promoting far transfer, in this study operationalized as real-life problem solving. Additional studies exploring methods for enhancing far transfer are warranted. Given the promising  $ES$ s on this study's far-transfer measure, researchers might consider marrying SBI with other approaches toward that end.

Before closing, we note that research assistants, not teachers, delivered classroom and tutoring SBI, with close monitoring of the fidelity of implementation. Therefore, this investigation qualifies as an efficacy study, not an effectiveness study. That is, findings demonstrate that SBI promotes superior outcomes at the classroom and tutoring prevention tiers under conditions that ensure strong SBI implementation. The present study does not address issues of effectiveness, that is, whether SBI effects can be realized when implemented by teachers (for classroom SBI) and school-based support staff (for tutoring SBI), and does not address issues about how to scale up Hot Math SBI. Given the demonstration of Hot Math SBI efficacy in the present study, research focused on issues of effectiveness and scaling up appear warranted.

## References

- Boaler, J. (1993). Encouraging the transfer of "school" mathematics to the "real world" through the integration of process and content, context, and culture. *Educational Studies in Mathematics*, 25, 341–373.
- Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (pp. 61–100). Washington, DC: American Educational Research Association.
- Brown, A. L., Campione, J. C., Webber, L. S., & McGilly, K. (1992). Interactive learning environments: A new look at assessment and instruction. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative view of aptitude, achievement, and instruction* (pp. 37–75). Boston: Kluwer Academic.
- Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, 18(1), 32–42.
- Burton, G. M., & Maletsky, E. M. (1999). *Math advantage*. Orlando, FL: Harcourt, Brace, Jovanovich.
- Chen, Z. (1999). Schema induction in children's analogical problem solving. *Journal of Educational Psychology*, 91, 703–715.
- Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121–152.
- Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem solving transfer. *Journal of Educational Psychology*, 79, 347–362.
- Deci, E. L., & Chandler, C. L. (1986). The importance of motivation for the future of the LD field. *Journal of Learning Disabilities*, 19, 587–594.
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.
- De Corte, E., Verschaffel, L., & Eynde, P. O. (2000). Self-regulation: A characteristic and a goal of mathematics education. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 687–726). San Diego: Academic Press.
- Durnin, J. H., Perrone, A. E., & MacKay L. (1997). Teaching problem solving in elementary school mathematics. *Journal of Structural Learning and Intelligent Systems*, 13, 53–69.
- Fuchs, D., Compton, D. L., Fuchs, L. S., & Davis, G. C. (in press). Making "secondary intervention" work in a three-tier responsiveness-to-intervention model: Findings from the first-grade longitudinal study at the National Research Center on Learning Disabilities. *Reading and Writing: A Contemporary Journal*.
- Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology*, 97, 493–513.
- Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., et al. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology*, 98, 29–43.
- Fuchs, L. S., Fuchs, D., Finelli, R., Courey, S. J., & Hamlett, C. L. (2004). Expanding schema-based transfer instruction to help third graders solve real-life mathematical problems. *American Educational Research Journal*, 41, 419–445.
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., Phillips, N. B., Karns, K., & Dutka, S. (1997). Enhancing students' helping behavior during peer-mediated instruction with conceptual mathematical explanations. *Elementary School Journal*, 97, 223–250.
- Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C. L., Dutka, S., & Katzaroff, M. (2000). The importance of providing background information on the structure and scoring of performance assessments. *Applied Measurement in Education*, 13(1), 1–34.
- Fuchs, L. S., Fuchs, D., Prentice, K., Burch, M., Hamlett, C. L., Owen, R., et al. (2003a). Enhancing third-grade students' mathematical problem



- solving with self-regulated learning strategies. *Journal of Educational Psychology*, 95, 306–315.
- Fuchs, L. S., Fuchs, D., Prentice, K., Burch, M., Hamlett, C. L., Owen, R., et al. (2003b). Explicitly teaching for transfer: Effects on third-grade students' mathematical problem solving. *Journal of Educational Psychology*, 95, 293–304.
- Fuchs, L. S., Fuchs, D., Prentice, K., Hamlett, C. L., Finelli, R., & Courey, S. J. (2004). Enhancing mathematical problem solving among third-grade students with schema-based instruction. *Journal of Educational Psychology*, 96, 635–647.
- Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1990). *Curriculum-based math computation and concepts/applications*. Unpublished manuscript, Vanderbilt University, Nashville, TN.
- Fuchs, L. S., Seethaler, P. M., Powell, S. R., Fuchs, D., Hamlett, C. L., & Fletcher, J. M. (2008). Effects of preventative tutoring on the mathematical problem solving of third-grade students with math and reading difficulties. *Exceptional Children*, 74, 155–173.
- Gagne, R. M. (1968). Contributions of learning to human development. *Psychological Review*, 75, 177–191.
- Garber, J., & Seligman, M. E. P. (Eds.). (1980). *Human helplessness*. New York: Academic.
- Geary, D. C. (1990). A componential analysis of an early learning deficit in mathematics. *Journal of Experimental Child Psychology*, 49, 363–383.
- Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, 15, 1–38.
- Gilbert, J., Compton, D. C., Fuchs, D., Fuchs, L. S., & Schatschneider, C. (2007). *Substitute or supplement: Exploring the nature of Tier-2 intervention*. Unpublished manuscript, Vanderbilt University, Nashville, TN.
- Glaser, R. (1983). Education and thinking: The role of knowledge. *American Psychologist*, 39, 93–104.
- Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- Greenes, C., Larson, M., Leiva, M. A., Shaw, J. M., Stiff, L., Vogeli, B. R., & Yeatts, K. (2005). *Houghton Mifflin math*. Boston: Houghton Mifflin.
- Gross-Tsur, V., Manor, O., & Shalev, R. S. (1996). Developmental dyscalculia: Prevalence and demographic features. *Developmental Medicine and Child Neurology*, 37, 906–914.
- Hanich, L. B., Jordan, N. C., Kaplan, D., & Dick, J. (2001). Performance across different areas of mathematical cognition in children with learning disabilities. *Journal of Educational Psychology*, 93, 615–626.
- Jitendra, A., DiPipi, C. M., & Perron-Jones, N. (2002). An exploratory study of schema-based word problem solving instruction for middle school students with learning disabilities: An emphasis on conceptual and procedural understanding. *Journal of Special Education*, 36, 23–38.
- Jitendra, A. K., Griffin, C. C., Haria, P., Leh, J., Adams, A., & Kaduvettor, A. (2007). A comparison of single and multiple strategy instruction on third-grade students' mathematical problem solving. *Journal of Educational Psychology*, 99, 115–127.
- Jitendra, A. K., Griffin, C. C., McGoey, K., Gardill, M. C., Bhat, P., & Riley, T. (1998). Effects of mathematical word problem solving by students at risk or with mild disabilities. *Journal of Educational Research*, 91, 345–355.
- Jitendra, A. K., & Hoff, K. (1996). The effects of schema-based instruction on the word-problem-solving performance of students with learning disabilities. *Journal of Learning Disabilities*, 29, 421–431.
- Jordan, N., & Hanich, L. (2000). Mathematical thinking in second-grade children with different forms of LD. *Journal of Learning Disabilities*, 33, 567–578.
- Jordan, N., & Montani, T. (1997). Cognitive arithmetic and problem solving: A comparison of children with specific and general mathematics difficulties. *Journal of Learning Disabilities*, 30, 624–634.
- Kansas State Board of Education. (1991). *Kansas quality performance accreditation*. Topeka, KS: Author.
- Larkin, J. H. (1989). What kind of knowledge transfers? In L. B. Resnick (Ed.), *Knowing, learning, and instruction* (pp. 283–305). Hillsdale, NJ: Erlbaum.
- Lester, F. K., & Garofalo, J. (1982, April). *Metacognitive aspects of elementary students' performance on arithmetic tasks*. Paper presented at the annual meeting of the American Educational Research Association, New York.
- Lewis, C., Hitch, G. J., & Walker, P. (1994). The prevalence of specific arithmetic difficulties and specific reading difficulties in 9- to 10-year-old boys and girls. *Journal of Child Psychology and Psychiatry*, 35, 283–292.
- Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenger, O. (2006). *SAS system for mixed models* (2nd ed.). Cary, NC: SAS Institute.
- Mawer, R., & Sweller, J. (1985). What do students learn while solving mathematics problems? *Journal of Educational Psychology*, 77, 272–284.
- Mayer, R. E. (1992). *Thinking, problem solving, cognition* (2nd ed.). New York: Freeman.
- Mayer, R. E., Quilici, J. L., & Moreno, R. (1999). What is learned in an after-school computer club? *Journal of Educational Computing Research*, 20, 223–235.
- McGrew, K. S., & Woodcock, R. W. (2001). *Technical manual. Woodcock-Johnson III*. Itasca, IL: Riverside.
- Prawat, R. S. (1992). Teachers' beliefs about teaching and learning: A constructivist perspective. *American Journal of Education*, 100, 354–395.
- Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology*, 88, 144–161.
- Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement, and instruction* (pp. 37–75). Boston: Kluwer Academic.
- Russell, R. L., & Ginsburg, H. P. (1984). Cognitive analysis of children's mathematical difficulties. *Cognition and Instruction*, 1, 217–244.
- Schoenfeld, A. H. (1992). Learning to think mathematically: Problem solving, metacognition, and sense-making in mathematics. In D. A. Grouws (Ed.), *Handbook of research on mathematics learning and teaching* (pp. 334–370). New York: Macmillan.
- Shalev, R. (2007). Prevalence of developmental dyscalculia. In D. B. Berch & M. M. M. Mazzocco (Eds.), *Why is math so hard for some children? The nature and origins of mathematical learning difficulties and disabilities* (pp. 49–60). Baltimore: Brookes.
- Silver, E. A., Branca, N., & Adams, V. (1980). Metacognition: The missing link in problem solving. In R. Karplus (Ed.), *Proceedings of the Fourth International Congress of Mathematics Education* (pp. 429–433). Boston: Birkhauser.
- Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, 2, 59–89.
- VanDerHeyden, A. M., & Witt, J. C. (2005). Quantifying context in assessment: Capturing the effect of base rates on teacher referral and a problem-solving model of identification. *School Psychology Review*, 23, 339–361.
- VanDerHeyden, A. M., Witt, J. C., & Gilbertson, D. A. (2007). Multi-year evaluation of the effects of a response to intervention (RTI) model on identification of children for special education. *Journal of School Psychology*, 45, 225–256.
- Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. *Learning Disabilities Research & Practice*, 18, 137–146.

- Vaughn, S., Linan-Thompson, S., & Hickman, P. (2003). Response to instruction as a means of identifying students with reading/learning disabilities. *Exceptional Children*, 69, 391-409.
- Vellutino, F., Scanlon, D. M., Sipay, E. R., Small, S. G., Pratt, A., Chen, R., et al. (1996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology*, 88, 601-638.
- Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence*. San Antonio, TX: Psychological Corporation.
- Woodcock, R. W. (1998). *Woodcock Reading Mastery Test-Revised*. Circle Pines, MN: American Guidance Service.
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III*. Itasca, IL: Riverside.
- Xin, P. X., Jitendra, A. K., & Deatline-Buchman, A. (2005). Effects of mathematical word problem-solving instruction on middle school students with learning problems. *Journal of Special Education*, 39, 181-192.
- Zimmerman, B. (1995). Self-regulation involves more than metacognition: A social cognitive perspective. *Educational Psychologist*, 30, 217-222.

Received September 27, 2007

Revision received February 14, 2008

Accepted February 17, 2008 ■

### Members of Underrepresented Groups: Reviewers for Journal Manuscripts Wanted

If you are interested in reviewing manuscripts for APA journals, the APA Publications and Communications Board would like to invite your participation. Manuscript reviewers are vital to the publications process. As a reviewer, you would gain valuable experience in publishing. The P&C Board is particularly interested in encouraging members of underrepresented groups to participate more in this process.

If you are interested in reviewing manuscripts, please write to the address below. Please note the following important points:

- To be selected as a reviewer, you must have published articles in peer-reviewed journals. The experience of publishing provides a reviewer with the basis for preparing a thorough, objective review.
- To be selected, it is critical to be a regular reader of the five to six empirical journals that are most central to the area or journal for which you would like to review. Current knowledge of recently published research provides a reviewer with the knowledge base to evaluate a new submission within the context of existing research.
- To select the appropriate reviewers for each manuscript, the editor needs detailed information. Please include with your letter your vita. In the letter, please identify which APA journal(s) you are interested in, and describe your area of expertise. Be as specific as possible. For example, "social psychology" is not sufficient—you would need to specify "social cognition" or "attitude change" as well.
- Reviewing a manuscript takes time (1-4 hours per manuscript reviewed). If you are selected to review a manuscript, be prepared to invest the necessary time to evaluate the manuscript thoroughly.

Write to Journals Office, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.



# Social Comparison and Big-Fish–Little-Pond Effects on Self-Concept and Other Self-Belief Constructs: Role of Generalized and Specific Others

Herbert W. Marsh  
University of Oxford

Ulrich Trautwein and Oliver Lüdtke  
Max Planck Institute for Human Development

Olaf Köller  
Humboldt University

Two studies integrate the big-fish–little-pond effect (BFLPE; negative effects of class-average achievement on academic self-concept, ASC), which is based upon educational psychological research, with related social psychological research that is based on social comparison theory. Critical distinctions are the nature of the social comparison processes that are based on generalized-other (class- or school-average) or individual (target comparison classmate) comparisons, and the nature of self-belief constructs that invoke normative (social comparison) or absolute frames of reference. In a large cross-national study (26 countries; 3,851 schools; 103,558 students), school-average ability negatively affected ASC but had little effect on 4 other self-belief constructs that did not invoke social comparison processes. In Study 2 (64 classes; 764 students), 2 sources of social comparison information (class-average achievement and achievement of an individually selected target comparison classmate) each had distinct, substantial negative effects on agency self-beliefs that invoked social comparison processes but not on metacognitive responses that did not invoke these processes.

**Keywords:** big-fish–little-pond effect, self-concept, social comparison theory, multilevel modeling, cross-national comparisons

The phenomenon of the self is widely accepted as an important universal aspect of being human and central to understanding the quality of human existence (Bandura, 2006; Branden, 1994; Bruner, 1996; Marsh & Craven, 2006; Pajares & Schunk, 2005). Positive self-beliefs are at the heart of the recent emphasis on a positive psychology. However, self-concept theory emphasizes that perceptions of the self cannot be adequately understood if the role of frames of reference is ignored. The same objective characteristics and accomplishments can lead to disparate self-concepts depending on the frames of reference or standards of comparison that individuals use to evaluate themselves, and these self-beliefs have important implications for future choice, performance, and behaviors.

Marsh and colleagues (see Marsh, 1991, 2007; Marsh & Craven, 2002; Marsh & Hau, 2003) proposed the big-fish–little-pond effect

(BFLPE) to encapsulate frame-of-reference effects in educational settings. In the BFLPE model, it is hypothesized that students compare their own academic ability with that of their classmates and use this social comparison information as one basis for forming their own academic self-concept (ASC). A negative BFLPE (a contrast effect) occurs when students have lower ASCs when they compare themselves to more able students and they have higher ASCs when they compare themselves with less able students. Consistent with theoretical predictions and a growing emphasis on the multidimensionality of self-concept, this research shows that the BFLPE is very specific to ASC; self-esteem and nonacademic components of self-concept are relatively unrelated to both individual- and group-average academic achievement.

## BFLPE: Effect Sizes, Social Comparison, and Item Wording

Herbert W. Marsh, Department of Educational Studies, University of Oxford, Oxford, England; Ulrich Trautwein and Oliver Lüdtke, Center for Educational Research, Max Planck Institute for Human Development, Berlin, Germany; Olaf Köller, Institute for Educational Progress, Humboldt University, Berlin, Germany.

Work on the present investigation was conducted, in part, while Herbert W. Marsh was a visiting scholar at the Center for Educational Research at the Max Planck Institute for Human Development in Berlin and was supported in part by the University of Oxford and the Max Planck Institute.

Correspondence concerning this article should be addressed to Herbert W. Marsh, Department of Educational Studies, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY, United Kingdom. E-mail: herb.marsh@education.ox.ac.uk

How well do these BFLPE findings, which are so robust in relation to ASC, generalize to other academic self-belief constructs such as agency, control, persistence, efficacy, metacognition and, specifically, self-efficacy? (For more discussion, see Bong & Skaalvik, 2003; Marsh, 1993; Skaalvik & Bong, 2003; Skaalvik & Rankin, 1995; Skaalvik & Skaalvik, 2002.) In particular, Marsh (1993) argued that frame-of-reference effects are directly implicated in ASC responses as individuals use the performances of others to establish frames of reference for evaluating their own performances. In contrast, little emphasis has been placed on the frames of reference that participants use to evaluate their performances in specific self-efficacy research. Bandura (1986), for

example, noted that self-esteem and self-concept—but not self-efficacy—are partly determined by “how well one’s behavior matches personal standards of worthiness” (p. 410). In specific self-efficacy judgments, the focus of assessment is on the individual’s capabilities in relation to the specific criterion items presented, and so the influence of frame of reference effects is minimized. Thus, for example, when students are shown example math test items and asked the probability of correctly answering such items, their responses are based on an absolute criterion that does not require them to compare their own performances with those of other students (also see Bong & Skaalvik, 2003).

For present purposes we distinguish between self-belief constructs, such as ASC, that (implicitly or explicitly) invoke normative comparisons and those, such as self-efficacy, that do not. The relevance of this distinction has been highlighted in a series of studies testing the internal/external frame of reference (I/E) model (Marsh, Walker, & Debus, 1991; also see Skaalvik & Rankin, 1995; Skaalvik & Skaalvik, 2002). According to the I/E model, students use an *external* (social comparison) frame of reference such as that posited in the BFLPE but also use an *internal* (ipsative-like) frame of reference in which students compare their own performance in one school subject (e.g., mathematics) with their own performance in other school subjects (e.g., English). Marsh et al. found clear support for frame-of-reference effects consistent with the internal comparison process for self-concept responses but not for self-efficacy responses. In exploring the implications of their results, Marsh et al. suggested that test-specific self-efficacy ratings severely truncated the operation of frame-of-reference effects and the evaluative responses to self-perceptions. For example, in judging the likelihood of correctly answering a test item on a math test, students have little need to compare their performance in one subject with their own performances in other subjects such as English (internal comparison) or with the performances of other students in mathematics (external comparison). Furthermore, a response of 80% likelihood of successfully answering a particular math problem does not indicate whether the student considers this to be a good or bad outcome. However, Marsh et al. also noted that some measures purporting to measure self-efficacy are based on stimuli likely to invoke social comparisons with other students (e.g., “I’m certain I can do an excellent job on assignments and tests,” where the term *excellent* might imply a comparison with the work of others). Hence, the empirically demonstrated distinction between self-concept and self-efficacy responses is likely to depend on the nature and wording of the items rather than on the label assigned to the construct.

A similar distinction can be made in terms of *postdictions* of performance, student judgments of their success following performance but before they know the outcome of their performance. Self-regulated learners are skillful at monitoring their learning process, including, for example, planning and self-efficacy prior to performance, attention control and self-regulation during performance, and appropriate self-reflection following performance (Zimmerman, 1998). Judgments of learning (Isaacson & Fujita, 2006; Nelson & Narens, 1994; Tobias & Everson, 2002) are an aspect of metacognitive knowledge monitoring that occurs during or after learning and are relevant to future planning and self-regulation. Thus, Isaacson and Fujita (2006; also see Hacker, Bol, Horgan, & Rakow, 2000) asked students to predict their test results

after completing the test but before they knew the test outcome—postdictions of performance. More accurate postdictions were associated with higher levels of achievement, and high-achieving students gave increasingly more accurate postdictions over time. Like self-efficacy ratings, postdicted performance can be assessed in relation to absolute standards of performance (e.g., number of test items correctly answered) rather than in relation to normative frames of reference that invoke a social comparison process.

There now seems to be clear empirical support for the normative and absolute distinction in the wording of items in different self-belief constructs and its influence upon the internal frame-of-reference effect—how a student’s self-beliefs in any one school subject are influenced by the same student’s performances in other school subjects. This previous research, however, has not extended the evaluation of this distinction to the external (social comparison) process posited in the BFLPE. Furthermore, although clearly supporting the distinction between specific self-efficacy and self-concept responses, this previous research has not, more fully, evaluated suggestions by Marsh et al. (1991) that this distinction has broader relevance to a range of self-belief items that vary in terms of this normative and absolute distinction. Hence, an overarching purpose of the present investigation is to explore the generalizability of the BFLPE to a variety of self-belief constructs associated with competence, agency, control, persistence, and test-performance expectations—as well as self-concept—and to relate variations in the size of the BFLPE in different constructs to this normative–absolute distinction.

#### Alternative Sources of Social Comparison Information: The Generalized Versus Specific Other

A major distinction between most social comparison theory (SCT) studies in social psychological research and BFLPE studies in educational psychology is the source of social comparison information. In the BFLPE, students are posited to use a generalized other as an implicit basis of comparison. This generalized other is operationalized as the mean performance level of other students in the same class or school. The process is implicit in that students are not explicitly instructed to make comparisons with other students. In social psychological research based on SCT, the focus is usually on a single target comparison person whom the participant explicitly chooses (e.g., Blanton, Buunk, Gibbons, & Kuyper, 1999; Diener & Fujita, 1997; Huguet, Dumas, Monteil, & Genestoux, 2001; Seaton et al., in press; Suls & Wheeler, 2000; Wheeler, 1966; Wood, 1989, 1996). The process is explicit in that the participant is explicitly instructed to choose a target comparison person. Implicit in this process is the assumption that students may use any one or a combination of a wide variety of different strategies in the selection of a target person (e.g., choosing someone more able because the student wants to identify and be more like the target person; choosing a target who is more able in order to learn how to perform the task better; choosing a target person who is less able in order for the student to look good and to protect the student’s self-concept; choosing a target person who is of similar ability to evaluate how well the student did in a particular situation; etc., also see Wood, 1989). In order to allow students as much flexibility as possible in choosing the target person, researchers typically give students little or no instruction on how to choose the target person. There is surprisingly little research jux-



taping results based on implicit comparisons with a generalized other emphasized in BFLPE studies and explicit comparisons based on a specific target comparison person emphasized in SCT research, even though the approaches are based on apparently similar theoretical rationales.

Because Festinger's early research that led to his development of SCT emphasized group processes, it is not surprising that he also emphasized this notion of a generalized other as a basis of normative comparisons between the self and a group—how individuals use groups to evaluate their abilities and opinions (Festinger, 1954; also see Suls & Wheeler, 2000). What we find surprising is that this historically important construct of generalized other has had so little emphasis in more recent SCT research that has focused more on variations of what became known as the rank-order paradigm, the choice of specific target individuals as a basis of comparison, the comparison of upward and downward comparison strategies, and how these strategies satisfy competing needs.

In BFLPE research, positive effects of attending schools in which the school-average ability is high (e.g., assimilation effects associated with reflected glory or pride associated with attending an academically selective school that leads to higher ASCs) has been elusive (see Lüdtke, Köller, Marsh, & Trautwein, 2005; Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006). Even when positive assimilation effects have been found, they are overshadowed by a counterbalancing negative contrast effect so that the net effect of attending academically selective schools on ASC is negative (Marsh, Kong, & Hau, 2000). Similarly, based on their review of SCT research, Buckingham and Alicke (2002) noted that while factors such as task relevance, similarity to the comparison target, cognitive load, and perceived control may be relevant,

people generally evaluate themselves more positively when the comparison information reflects favorably (i.e., following downward comparisons) rather than unfavorably (i.e., following upward comparisons) on their characteristics and abilities, especially when they receive direct feedback regarding their own and others' behavioral or performance outcomes. (p. 1117)

However, unlike BFLPE research in which there is consistent support for the negative effects of upward comparisons, the theoretical predictions and empirical results are not so clear in SCT studies. In particular, when participants are able to choose a target person with whom to compare, upward comparisons sometimes result in assimilation rather than contrast, leading Major, Testa, and Bylsma (Major, Testa, and Bylsma, 1991, p. 238) and many others to describe social comparisons as a "double-edged sword." An important focus of much of this SCT research has been the strategies that individuals use to select comparison targets (e.g., upward and downward comparison strategies) to maximize competing needs. Thus, upward evaluations might provide a basis of identification with more accomplished target persons, even though such target persons are likely to provide a more demanding basis of comparison for self-evaluations, leading to feelings of inferiority. Nevertheless, when asked to choose target persons with whom to compare themselves, participants typically choose targets who are similar or slightly better than themselves (i.e., upward rather than downward; see Blanton et al., 1999; Huguet et al., 2001; Suls & Wheeler, 2000; Wheeler, 1966).

Whereas the emphasis on generalized others in BFLPE studies may be a reasonable assumption within an imposed social comparison paradigm in educational settings (Diener & Fujita, 1997), more research is needed to test the generalizability of this effect over different sources of social comparison information such as that provided by individual target comparison persons chosen by students in free-choice situations. Furthermore, the uses of generalized and specific comparison targets are not mutually exclusive. Individuals might simultaneously evaluate their performances in relation to both the performances of specific target individuals selected in ways that have been considered in SCT research and in relation to some generalized performance based on a group-average performance, as posited in BFLPE research.

Reuman (1989) conducted one of the few studies to juxtapose a generalized and specific other. Consistent with the BFLPE, he found that ability grouping (i.e., assigning students to low-, medium-, or high-ability classes rather than to more heterogeneous, mixed-ability classes) produced lower ASCs for high-ability children and higher ASCs for low-ability children. When students were asked to compare their test scores with those of a classmate and whether the selected classmate was more or less able, ability grouping was associated with systematic differences in the perceived ability of the comparison classmate. Children from high-ability classes were more likely to select classmates with higher abilities than their own (upward comparison), and children from low-ability classes were more likely to select classmates with lower abilities than their own (downward comparisons). However, Reuman did not evaluate whether the direction of comparison affected ASC and whether this effect of comparison direction was independent of the average ability levels of the different classes.

It is also relevant to note that the distinction between normative and absolute frames of reference that is one focus of the present investigation has not been widely considered or integrated into SCT. However, Smith and Arnelsson (2000) made a similar distinction in relation to the possible referents of ability appraisal in SCT: ability relative to absolute standards demanded of a task (e.g., how quickly I can run 1 km—an absolute judgment that does not require relativistic responses); ability relative to specific others (as emphasized in SCT); and ability relative to a reference group (a response in relation to a generalized other as emphasized in BFLPE studies). Smith and Arnelsson noted that in order to most accurately estimate ability relative to a group, participants should use an aggregate index of group performance (as in the class-average achievement in BFLPE studies), and this preference for class-average performance was demonstrated by Suls and Tesch (1978).

### The Present Investigation

In the present investigation we conducted two studies to test the generalizability of the BFLPE in relation to different self-belief constructs posited to invoke different frames of reference and to different sources of social comparison information (also see Trautwein & Lüdtke, 2005). Study 1 was based on a large cross-national sample in which we evaluated the effects of school-average achievement on ASC and a variety of other academic self-beliefs that varied in relation to which frame of reference they invoked: normative or absolute. Study 2 was based on a large



representative sample of German high school students and included self-belief constructs that varied in terms of the absolute and normative distinction that was the focus of Study 1, but also in terms of different sources of social comparison information based on generalized others (class-average achievement) and specific others (selection of a specific target comparison classmate). In both studies we evaluated the generalizability of the BFLPE based on ASC to other academic self-belief constructs and tested our prediction that BFLPEs would be systematically larger for constructs that invoke a normative or social comparison frame of reference than for constructs that invoke an absolute frame of reference.

In Study 2, we sought to integrate the distinctions made in BFLPE and SCT studies in relation to sources of social comparison information. This is apparently the first BFLPE study to simultaneously evaluate the combined effects of upward comparisons based on a generalized other (as operationalized by class-average achievement in BFLPE studies) and a specific other (as operationalized by free choice of a specific target comparison person in SCT studies). Although we posited that the effect of upward comparisons in relation to both sources of social comparison information would be negative (negative effects of class-average achievement and upward comparison choice), the theoretical and empirical support for these predictions is much stronger for generalized others (from BFLPE studies) than for specific others (from SCT studies). Of particular relevance are the questions of the extent to which the two sources of social comparison make independent contributions to self-belief constructs when both are considered simultaneously and whether the size (or even the direction of assimilation instead of contrast) of either changes substantially when the other is considered.

### Study 1

Study 1 was a large cross-cultural test of the generalizability of the BFLPE based on nationally representative samples of students from 26 countries. We evaluated the generalizability of the BFLPE based on ASCs (Marsh & Hau, 2003) across a range of other self-belief factors. For ASC, consistent with the BFLPE, we predicted that (a) the effect of individual student achievement would be positive and not vary substantially from country to country and (b) the effect of school-average achievement would be negative and not vary substantially from country to country. The central research question in Study 1 was how well these results based on ASC generalize to four additional self-belief constructs: general self-efficacy, control expectations, control strategies, and effort persistence. Each of these self-belief constructs has a rich theoretical and empirical basis that led to their selection by expert panels of the Organisation for Economic Cooperation and Development (OECD) as being among the best, most useful, and psychometrically strongest constructs in educational psychology. (For a detailed discussion of these constructs and the basis of their selection, see Marsh, Hau, Artelt, Baumert, & Peschar, 2006; see wording of the items in Appendix A.) However, in the present investigation, we focused particularly on our observation that each of these constructs is based on responses to items that invoke a more absolute frame of reference than the ASC responses, leading us to predict that the BFLPE will be smaller for these constructs than for ASC.

However, because of the nature and wording of these other self-belief constructs (Appendix A), we anticipated that there would be some normative frame-of-reference effects resulting in a negative effect of school-average ability. Thus, for example, in order to respond to the general self-efficacy item "I'm confident I can do an excellent job on assignments and tests," students would be likely to use the other students in their educational context as a basis of comparison in forming judgments about what constitutes excellent performance, their level of confidence (i.e., "How confident am I relative to other students?"), and their underlying ability that was a probable basis of their confidence. Hence, this general self-efficacy item is different from specific self-efficacy items considered in other research (see earlier discussion of Marsh et al., 1991) in which the criterion of success was explicit in the item and not dependent on comparisons with an external, normative standard (e.g., other students in the class or school). Similar observations apply to items from the conceptually similar control expectation scale (e.g., "If I decide not to get any bad grades, I can really do it"). The implied use of other students as a basis of comparison is less clear for items from the effort persistence ("When studying, I keep working even if the material is difficult") and control strategy ("When I study, I force myself to check to see if I remember what I have learned") in that these items do not necessarily require comparisons with other students.

### Method

*Data source and sample.* Study 1 was based on the Programme for International Student Assessment (PISA) database compiled by the OECD that consists of responses by nationally representative samples of approximately 4,000 15-year-olds collected in 32 countries in the year 2000 (for a description of the database and variables, see OECD, 2001a, 2001b, 2003; also see Marsh, Hau, et al., 2006). The PISA database was collected in response to the need for internationally comparable evidence of student performance and related competencies within a common framework that was internationally agreed upon. Selection of the measures was made on the basis of advice from substantive and statistical expert panels and the results from extensive pilot studies. In many cases (see Marsh, Hau, et al., 2006) the selection of items used to represent a particular scale was a subset of the items from a widely used instrument. Thus, for example, the ASC items (Appendix A) selected for inclusion in the PISA study were a subset of the 10 items in the Self Description Questionnaire II (Marsh, 1990, 1992), selected on the basis of those with the best psychometric properties in previous research and on the subsequent pilot studies conducted by the OECD (see Marsh, Hau, et al., 2006). Substantial efforts and resources were devoted to achieving cultural and linguistic breadth in the assessment materials, stringent quality assurance mechanisms were applied in the translation of materials into different languages, and data were collected under independently supervised test conditions.

Paper-and-pencil assessments consisted of a combination of multiple-choice items and written responses. While all students completed some reading assessment items, which were the focus of the 2000 data collection, only random samples of students completed mathematics or science assessments. In addition, countries were given the option of collecting materials on the Student Approaches to Learning instrument (Marsh, Hau, et al., 2006) that



included all self-belief scales that were the focus of the present investigation. Twenty-six of 32 countries administered this additional survey. Evaluations of responses to these scales based on this database demonstrated good psychometric properties (a well-defined factor structure based on multiple-group confirmatory factor analyses across the 26 countries, reliability, and validity in relation to academic achievement; see Marsh, Hau, et al., 2006). Coefficient alpha estimates of reliability for the five scales considered here are (see Appendix A): ASC (.78), self-efficacy (.75), control expectations (.75), control strategies (.76), and effort-persistence (.78).

Study 1 was based on students who completed the Student Approaches to Learning instrument and standardized academic achievement tests that were developed specifically for PISA. In preliminary analyses, confirmatory factor analyses of responses to items representing these five scales indicated that the five a priori factors were well-defined in that factor loadings were consistently high and the goodness of fit for the a priori factor structure provided an excellent fit to the data (see Appendix B).

Due to the nature of the PISA project, nine versions of the achievement tests were administered that contained different combinations of verbal, mathematics, and science test items. Owing to the focus of the PISA study, all students completed at least some verbal test items, whereas only about half of the students completed math and science tests. Using test-equating procedures based on item response theory, scores based on each version of the various achievement tests were put onto a comparable scale. For present purposes, a total achievement test score was obtained by taking an average score for all the achievement test scores available for each student. As recommended in the database documentation (OECD, 2001a, 2001b), analyses were conducted using sample weights to obtain unbiased estimates of population parameters. For purposes of the present investigation, the effective sample size for each country was set to be equal to the number of cases for that country prior to weighting, so that the weighted sample size was the same as the unweighted sample size (i.e., the average weight across all cases was 1.0; see Marsh & Hau, 2003).

*Statistical analysis.* In most studies conducted in school settings, individual student characteristics are confounded with classroom or school characteristics because individuals are not randomly assigned to groups. This clustering effect introduces problems related to appropriate levels of analysis, aggregation bias, and heterogeneity of regression (Goldstein, 2003; Raudenbush & Bryk, 2002). In the present investigation, for example, the meaning of a variable at the student level does not necessarily bear any straightforward relation to its meaning at the school level. The negative BFLPE is a dramatic example of this problem—achievement at the individual student level is positively related to ASC, whereas achievement at the school level is negatively related to ASC. The juxtaposition of the effects of individual and school-average achievement is inherently a multilevel issue that cannot be represented adequately at either the individual or the school level. A detailed presentation of multilevel modeling (also known as hierarchical linear modeling) is beyond the scope of the present investigation and is available elsewhere (e.g., Goldstein, 2003; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). All models reported here are random-intercept models estimated by iterative generalized least squares (using MLwiN, version 2.2; Rasbash, Steele, Browne, & Prosser, 2004). Consistent with the PISA de-

sign, we considered a three-level multilevel model in which students (Level 1) were nested within schools (Level 2) and schools were nested within countries (Level 3). In order to enhance interpretations of the results, we standardized (*z*-scored) all variables to have  $M = 0$ ,  $SD = 1$  across the entire sample (see Marsh & Rowe, 1996; also see Aiken & West, 1991; Raudenbush & Bryk, 2002). School-average measures of achievement were determined by taking the average of achievement scores for students in each school (but not restandardizing these scores so that individual student and school-average achievement scores were in the same metric).

In Study 1, individual student achievement and school-average achievement were related to the ASC and other self-belief constructs. We estimated fixed and random effects associated with the variables of interest (see Table 1). The fixed effects consisted of the main effects of individual student achievement and school-average achievement and a constant term (which is approximately zero because of the use of standardized variables and, thus, is of little practical interest). Random effects consisted of intercept terms associated with each level of the model, indicating the extent to which intercepts varied from country to country (Level 3), from school to school (Level 2), and from student to student (Level 1). These terms reflect the amount of unexplained (residual) variance at each of the three levels. In addition, the effects of individual achievement were allowed to be random at both the school and the country level, providing an indication of how much the effect of individual achievement on ASC varied from country to country and from school to school. Finally, the effects of school-average achievement were also allowed to be random at the country level, providing an indication of how much the effect of school-average achievement on ASC varies from country to country.

## Results and Discussion

In Model M1 (Table 1), we tested the basic BFLPE for ASC. Consistent with previous research (Marsh & Hau, 2003), there was a substantial positive effect (.38) of individual student achievement on ASC. Students who scored 1 *SD* above the mean of the academic achievement test on average scored .38 *SD* above the mean of the ASC scale. Of particular relevance to tests of the BFLPE, there was a substantial negative effect (−.20) of school-average achievement on ASC. On average, students attending schools in which the average achievement level was 1 *SD* above the mean achievement level (in the metric of the individual student achievement test scores) scored .2 *SD* below the mean of the ASC scale. Although the effects of individual and school-average achievement on ASC varied significantly from country to country, due in part to the extremely large sample size, this variation was very small (both residual variance components were .01). In summary, the results provide strong support for the generalizability of the BFLPE across the 26 countries included in this OECD database.

In Models M2–M5 (Table 1) we tested the generalizability of the BFLPE found with ASC to four other self-belief constructs. Consistent with a priori predictions, the negative effects of school-average achievements (the BFLPE) were systematically less negative for all four of these constructs (−.08 to .03) than that observed with ASC (−.20). In fact, whereas the negative effects of school-average achievement were small but statistically significant ( $p < .05$ ) and negative for self-efficacy (−.08) and control expectations

( $-.06$ ), the effects were not statistically significant for control strategies ( $.03$ ) and effort persistence ( $-.01$ ). These results clearly support predictions that the BFLPE would be smaller for these other self-belief constructs than for ASC.

Although differences among the sizes of the effects of the school-average achievement for the four additional self-belief constructs were small, it is interesting to speculate on the nature of these differences. Previous research (Marsh et al., 1991) found that frame-of-reference effects (based on a given student's performance in other school subjects) were much smaller for specific self-efficacy measures than for self-concept. Although we know of no direct comparisons of the generalizability of this distinction to social comparison frame-of-reference effects posited in the BFLPE, it might be expected that the negative effect of school-average achievement would be less negative for efficacy than for other self-belief scales. However, there are reasons why this expectation was not warranted in the present study. In particular, this previous research was based on test-specific self-efficacy measures in which students were asked to estimate the probability of correctly answering items like those shown to them—a task that did not require them to compare their relative abilities in different school subjects or, relevant to the present investigation, to compare their own ability levels with those of their classmates. In this respect, the generalized self-efficacy measure from the PISA study was quite different.

Indeed, as suggested earlier, the self-efficacy and control expectation items (see Appendix A) both might invoke students to make comparisons with other students in forming their response, whereas items from the control strategies and effort persistence scales apparently do not. This suggests that the lack of frame-of-reference effects associated with different self-belief constructs might be a function of the precise nature of the items used to measure each construct as well as inherent differences in content

of the constructs. Whereas the efficacy responses were apparently less subject to frame-of-reference effects than were the self-concept responses, it is also relevant to note that academic achievement was systematically more strongly related to ASC ( $.38$ ) than to any of the other self-belief constructs, and that the two self-belief constructs that did have significant frame-of-reference effects were more strongly related to achievement ( $.30$ , self-efficacy;  $.27$ , control expectations) than were the other two for which frame-of-reference effects were not significant ( $.19$ , control strategies;  $.15$ , effort persistence). Hence, not surprisingly, social comparison information typically provides a useful basis for forming accurate self-appraisals. In summary, the results of Study 1 clearly support a priori predictions about differences in the sizes of the BFLPE for ASC and the other self-belief constructs.

## Study 2

In Study 2 we extended findings from Study 1 and previous research in two ways.

1. Predictions about the lack of frame-of-reference effects in Study 2 were based on a metacognitive measure, math-test-specific performance postdictions (MTSPP), in which students estimated the number of items that they had correctly answered immediately after taking a test. While the results of Study 1 clearly demonstrated that the negative effects associated with the BFLPE were systematically smaller for a variety of self-belief constructs than for ASC, there was no test-specific measure in which measures were based on such a clearly absolute scale as the number of questions students predicted that they had correctly answered. (Note that performance expectations used here clearly differed from the traditional test-specific self-efficacy measures like those used by Marsh et al., 1991, in that expectations were in relation to a test already completed. However, the two measures are similar in

Table 1  
*Study 1: Generalizability of the Big-Fish–Little-Pond Effect Across Five Constructs*

Variables	M1 Academic Self	M2 Self-Efficacy	M3 Control Expectation	M4 Control Strategies	M5 Effort Persistence
	<i>b</i> ( <i>SE</i> )	<i>b</i> ( <i>SE</i> )	<i>b</i> ( <i>SE</i> )	<i>b</i> ( <i>SE</i> )	<i>b</i> ( <i>SE</i> )
Fixed effects					
Level 1: Individual Achievement	.38 (.024)*	.30 (.018)*	.27 (.020)*	.19 (.014)*	.15 (.017)*
Level 2: School-Avg Achievement	-.20 (.019)*	-.08 (.023)*	-.06 (.022)*	.03 (.020)*	-.01 (.021)
Residual variance components					
Level 3: Country					
Intercept	.09 (.050)	.05 (.023)*	.06 (.021)*	.09 (.026)*	.05 (.019)*
Individual Achievement	.01 (.004)*	.01 (.002)*	.01 (.002)*	.01 (.001)*	.01 (.002)*
School-Avg Achievement	.01 (.002)*	.01 (.005)*	.01 (.004)*	.01 (.003)*	.01 (.004)*
Level 2: School					
Intercept	.02 (.002)*	.02 (.003)*	.02 (.002)*	.02 (.003)*	.02 (.003)*
Individual Achievement	.01 (.001)*	.01 (.001)*	.01 (.001)*	.01 (.001)*	.01 (.001)*
Level 1					
Residual	.78 (.027)*	.87 (.027)*	.89 (.024)*	.86 (.025)*	.82 (.025)*

Note. All outcome and predictor variables were standardized ( $M = 0$ ,  $SD = 1$ ) at the individual student level. In each analysis, individual achievement and school-average achievement were related to one of the five self-report factors. All parameter estimates are statistically significant when they differ from zero by more than two standard errors (*SEs*).

\*  $p < .05$



that the response scale—the number of correctly answered items—is clearly an absolute scale rather than a normative scale that invokes social comparisons).

Although previous research has demonstrated the relevance of this absolute–normative comparison distinction in relation to frame-of-reference effects based on the juxtaposition of performances in different school subjects by the same student (an internal comparison; e.g., Marsh et al., 1991), we sought to determine whether these conclusions generalize to fundamentally different frame-of-reference effects based on the juxtaposition between performance by a particular student and those of his or her classmates (an external comparison). In particular, we were not aware of tests of this distinction in any previous BFLPE or SCT studies. In addition to this MTSP measure we also evaluated the BFLPE in relation to a measure of math agency that is based on generalized items (e.g., “I’m pretty smart in things that concern mathematics”) that invoke social comparison processes like those self-concept responses. On the basis of the nature of these two constructs, previous research, and the results of Study 1, we predicted that the BFLPE based on class-average math achievement would be systematically more negative for math agency than for MTSP, and that these frame-of-reference effects would be much smaller (or nonsignificant) for MTSP.

2. We compared results based on a generalized other (operationalized as class-average achievement as in BFLPE studies) and a specific other (operationalized as direction of comparison with a freely chosen target comparison classmate as in SCT studies) separately. We also examined the combined effects of both these sources of social comparison information. While contrast effects have consistently been found in BFLPE studies in which comparisons are implicit, the results have not been so consistent in other social comparison research in which participants have been explicitly instructed to choose a target person and have had considerable freedom in the strategies that they have used to make this choice (e.g., Buunk & Gibbons, 2000; Diener & Fujita, 1997). Perhaps reflecting this distinction, Stapel and Suls (2004) noted that implicit comparisons result in contrast effects (negative effects of upward comparison), but explicit comparisons can produce assimilation effects (positive effects of upward comparison). This distinction is apparently relevant in that comparison with a generalized other or class average in BFLPE research is implicit, whereas SCT studies are typically based on explicit comparisons. Hence this methodological distinction between comparisons involving (implicit) generalized others and an (explicit) chosen target comparison classmate is theoretically important. We expect both sources of social comparison information (class-average achievement and upward comparisons) to have negative effects on the more generalized measure of agency, consistent with the BFLPE, but the basis for predictions for the chosen target comparison classmate has less empirical support from previous research. However, we also predict that both sources of information will have less negative (or no) effect on MTSP.

## Method

*Sample.* The sample consisted of 736 students (49.8% girls) who had a mean age of 13.8 years ( $SD = .67$ ) enrolled in a representative sample of 64 mathematics classes from one large German state. The sample was representative of the state in terms

of school type: Hauptschule, 22.7%; Realschule, 24.6%; Gymnasium, 36.3%; and Comprehensive, 16.4%. Data were collected in the second half of the seventh grade. Because in this state students change to secondary school after Grade 4 and almost all students remain in intact classes throughout the secondary level, these students had already been taught mathematics in their secondary school classroom for more than 2 years prior to the time of the study. The test materials were administered by trained research assistants. The data used in Study 2 were part of a larger study on historic changes in school achievement with a total of 1,452 students. In all classes, the students were randomly allocated to two different testing conditions. Only half of the students were asked to answer the comparison task (see below), and these 736 students considered here constituted a random sample of the total group.

*Mathematics achievement.* The items constituting the math achievement test were taken from previous national and international studies, in particular the First and Second International Mathematics Study (Husén, 1967; Robitaille & Garden, 1989) and an investigation conducted at the Max Planck Institute for Human Development (cf. Baumert, Roeder, Sang, & Schmitz, 1986). Various content areas—arithmetic, geometry, elementary statistics, and algebra—were covered, and the curricular validity of all items had been assessed beforehand by curriculum experts. The reliability of the 46-item test was .85 (using the Kuder–Richardson formula 20).

*Self-belief scales.* Two measures of self-beliefs were used. To assess MTSP, students were asked to estimate the number of items on the mathematics test that they had solved correctly after they had completed the test (i.e., a postdiction). To assess math agency, we adapted three items from the Agency Beliefs: Ability Scale from the Control, Agency, and Means–Ends Interview (CAMI; T. D. Little, Oettingen, Stetsenko, & Baltes, 1995; see Skinner, Zimmer-Gembeck, & Connell, 1998) to the mathematics domain using a 4-point (*disagree–agree*) response scale (see Appendix A for wording of the items). The coefficient alpha estimate of reliability for agency responses based on results from the present investigation was .68.

*Comparison target choice and comparison direction.* It is important to emphasize that both the students’ freedom of choice in selecting a target comparison classmate and the direction of the comparison were critical components of this construct. After having finished their mathematics tests and postdicting the number of correctly answered items, students were instructed to nominate the student whose test booklet they would like to compare with their results. Following the procedure typically used in this area of research (e.g., Reuman, 1989; Swallow & Kuiper, 1992; also see Wood, 1989), students were not instructed as to how to make this selection or given any specific criteria to use in this decision. Thus, as would be the case outside of an experimental setting, students had complete freedom in choosing their comparison target. In particular, they were free to choose a target comparison classmate who was more able (upward comparison), less able (downward comparison), or equally able. Because this is the typical approach used in social psychological research (see Wood, 1996), it was important for us to use a similar procedure in the present investigation in which we sought to juxtapose these results with those based on the class-average achievement measure used in BFLPE.

After choosing a target comparison classmate, students were then asked to compare their typical mathematics performance with that of their chosen target comparison classmate. Students were asked to indicate "Is this student in mathematics (a) better than you? (b) not as good as you? (c) similar in achievement level?" A majority of students chose target comparison classmates perceived as better (55%), whereas 30% and 15% of the students chose target comparison students perceived to be not as good as and similar to them. In this respect the results were like those found in other social comparison research (see Suls & Wheeler, 2000) in that there was a tendency toward upward comparison (selecting a more able student), although there were also substantial numbers who chose comparison targets who were of similar or lower ability.

Three scores were constructed to represent these responses: upward comparison (upward = 1, else = 0), downward comparison (downward = 1, else = 0), and upward-downward comparison (upward = 1, equal = 0, downward = -1). Because the third score was completely determined by the first two (i.e., they were completely multicollinear), the purpose of using these alternative representations was to determine which one or combination of two scores provided the best and most parsimonious representation of this variable.

**Statistical analyses.** The statistical analyses conducted in Study 2 were similar to those in Study 1. Specifically, we conducted a multilevel analysis to assess the effects of individual student math achievement, class-average math achievement, and comparison direction on MTSPPE and math agency. As in Study 1, in order to enhance the interpretability of the effects, we standardized all independent and dependent variables ( $M = 0$ ,  $SD = 1$ ). Academic achievement was aggregated at the class level to form class-average achievement based on standardized test scores of individual students but was not restandardized so that class-average achievement was measured on the same metric as individual student achievement tests. All reported models are random-intercept models estimated by iterative generalized least squares (using MLwiN, version 2.2; Rasbash et al., 2004). Missing data represent a potentially serious methodological problem in many empirical studies. Whereas the amount of missing data (6%)

was not large, there is growing recognition of the inappropriateness of traditional approaches to missing data such as listwise and pairwise deletion of missing data (e.g., Enders, 2006; R. J. A. Little & Rubin, 1987; Schafer, 1997). In Study 2 we used NORM software (version 2.03, see Schafer & Graham, 2002) to impute missing values. The NORM software uses an expectation-maximization approach to estimate the variance-covariance matrix on which basis missing values are imputed.

### Results and Discussion

**Math agency.** The results of Model 1 for agency (Table 2) provide clear support of the typical BFLPE. In particular, there was a substantial positive predictive effect of individual student achievement and a substantial negative predictive effect of class-average achievement. In Model 2 for agency (Table 2) we tested the same model with the upward comparison variable instead of class-average achievement. Results based on this alternative source of social comparison information gave a similar pattern of results. In particular, the predictive effect of individual student achievement was positive, whereas the predictive effect of selecting a more able student as the target of comparison (upward comparison in Table 2) was negative. Finally, in Model 3 for agency (Table 2) we evaluated the combined predictive effects of both sources of social comparison information—school-average achievement (generalized other) and upward comparisons (specific other). Although the negative predictive effects of each of these sources of social comparison information were diminished somewhat—compared with models in which each was considered separately—the negative predictive effects of both class-average achievement and upward social comparison were highly significant ( $p < .001$ ) and substantially meaningful. In summary, this first set of models is consistent with a priori predictions in that the BFLPE was replicated for math agency responses based on both class-average achievement (the typical basis of social comparison information based on a generalized other in BFLPE studies) and upward comparison (a typical basis of social comparison informa-

Table 2  
Study 2: Generalizability of the Big-Fish-Little-Pond Effect Across Different Constructs and Different Sources of Comparison Information

Variables	Math Agency			Math Test Performance Expectations (MTSPPE)		
	M1 Agency	M2 Agency	M3 Agency	M1 MTSPPE	M2 MTSPPE	M3 MTSPPE
	<i>b</i> ( <i>SE</i> )	<i>b</i> ( <i>SE</i> )	<i>b</i> ( <i>SE</i> )	<i>b</i> ( <i>SE</i> )	<i>b</i> ( <i>SE</i> )	<i>b</i> ( <i>SE</i> )
Fixed effects						
Level-1: Individual Achievement	.44 (.056)*	.18 (.038)*	.36 (.057)*	.31 (.056)*	.27 (.038)*	.26 (.058)*
Level-1: Upward Comparison		-.25 (.035)*	-.20 (.036)*		-.12 (.035)*	-.12 (.036)*
Level 2: Class-avg Achievement	-.44 (.073)*		-.32 (.075)*	-.05 (.076)		.02 (.079)
Residual variance components						
Intercept	.01 (.015)*	.02 (.016)	.01 (.015)	.02 (.018)	.02 (.018)	.02 (.018)
Residual	.90 (.049)*	.87 (.048)*	.86 (.047)*	.90 (.049)*	.89 (.048)*	.88 (.048)*

Note. All outcome and predictor variables were standardized ( $M = 0$ ,  $SD = 1$ ) at the individual student level. All parameter estimates are statistically significant when they differ from zero by more than two standard errors (*SEs*).

\*  $p < .05$



tion based on comparison with a selected target person in social comparison studies). Furthermore, when both of these sources of social comparison information were considered simultaneously in a single model, each made substantial, unique contributions.

**MTSPP.** A parallel set of models was tested with MTSPP. Consistent with a priori predictions and the nature of the MTSPP task, the predictive effect of class-average achievement was substantially diminished. In fact, the predictive effect of class-average achievement ( $-.05$ ) was not statistically significant ( $p > .05$ ) despite the large sample size (see Model M1 for MTSPP in Table 2). In Model M2 for MTSPP, we tested a parallel model with the second source of social comparison information (upward comparison rather than class-average achievement). Here the predictive effect of upward comparison was significantly negative, although the size of the predictive effect was modestly less than its corresponding predictive effect on math agency ( $-.12$  for MTSPP,  $-.25$  for math agency). Hence, the difference in the negative predictive effects of class-average ability on the two outcome variables ( $-.44$  for agency vs.  $-.05$  for MTSPP) was substantially larger than for MTSPP. Finally, in Model M3 for MTSPP we considered, simultaneously, both sources of social comparison information. In this model, the predictive effect of class-average achievement was again not statistically significant, whereas the statistically significant negative predictive effect of upward comparison was almost identical to that in Model M2 for MTSPP. In summary, this second set of models is also consistent with a priori predictions in that the BFLPE was substantially diminished for MTSPP responses compared with those based on math agency. Whereas the predictive effect of upward comparison MTSPP was diminished, the predictive effect of class-average achievement was not even statistically significant.

**Supplemental analyses.** In supplemental analyses we explored several variations of these models based on both math agency and MTSPP responses. First, we explored the extent of generality of the predictive effects across different classes by allowing the individual student Level 1 variables (individual achievement and upward comparison) to be random in each of the various models. However, because the resulting variance components were not statistically significant (indicating a lack of class-to-class variation in the predictive effects), these supplemental analyses were not considered further. Nevertheless, these results are important, providing an extremely powerful test of and support for the generalizability of the results across the 64 math classes considered in Study 2.

Second, we explored the implications of considering alternative representations of the direction of comparison based on the target comparison classmate. The downward comparison variable (choosing a target person perceived to be less able) did not have statistically significant predictive effects for any of the models when considered in isolation (instead of upward comparison in models shown in Table 2) or when considered in combination with either of the other direction of comparison variables (upward comparison or up-down comparison). Whereas the up-down comparison variable (downward =  $-1$ , equal =  $0$ , upward =  $+1$ ) also had the expected negative predictive effect in each of the relevant models, its predictive effect was smaller than the upward comparison variable presented here. Based on these supplemental results, we present

only results based on the upward comparison that was the primary focus of the present investigation. However, these supplemental results are substantively important, indicating that predictive effects of the direction of comparison apparently are not symmetric. In particular, whereas there are negative predictive effects of upward comparison, there are apparently no positive predictive effects of downward comparison. Although beyond the scope of the present investigation (but see Aspinwall & Taylor, 1993; Diener & Fujita, 1997; Marsh, Tracey, & Craven, 2006), this is a finding that warrants further research.

## Overall Discussion and Implications

### *BFLPE: Importance of Normative and Absolute Frames of Reference*

Our study is apparently the first to provide clear support for the importance of distinguishing between normative and absolute frames of reference in the evaluation of the BFLPE. An overarching concern of both Studies 1 and 2 was how well the BFLPE generalizes to other academic self-belief constructs, particularly those that invoke a normative or social comparison frame of reference and those based on a more absolute frame of reference. The results of the present investigation are consistent with earlier research showing that the internal comparison process (using results in one school subject as a basis for formulating ASCs in another school subject) was evident only in self-concept responses that invoked a social comparison frame of reference but not in self-efficacy responses that invoked an absolute frame of reference. However, this distinction between normative and absolute frames of reference has apparently not been systematically evaluated in relation to the BFLPE, frame-of-reference effects based on comparisons with other students, and the broader field of social comparison research. More generally, the results make a potentially important contribution to better understanding the relevance of frame-of-reference effects to different academic self-belief constructs.

In Study 1 we compared the strength and generalizability of the BFLPE for self-concept with those for other self-belief constructs on the basis of nationally representative samples of 15-year-old students from 26 different countries. Results provided impressive support for the negative effect of school-average ability on ASC and the generalizability of this BFLPE across the different countries. Also consistent with a priori predictions, the BFLPEs were much weaker for each of four other self-belief constructs (generalized self-efficacy, control expectations, control strategies, and effort persistence). There were, however, differences in the size of the effects of school-average achievement in these other four self-belief constructs. Interestingly, the negative effect of school-average achievement was more negative for the construct actually labeled *self-efficacy* than for the other three self-belief constructs. These differences, however, were consistent with a priori predictions based on the nature of the wording of items—the extent to which they implicitly invoke social comparison processes in forming responses to them. This is consistent with our claim that it is the nature of the situational demands and the nature of the items used in self-belief measures that distinguishes this construct from the ASC construct in relation to frame-of-reference effects. Furthermore, this distinction between self-efficacy and self-concept



apparently does not hold—or holds to a lesser extent—for general self-efficacy measures for which the criterion of successful performance is not an explicit part of the item.

Predictions about the diminished role of social comparison information in the BFLPE were based in part on results for test-specific measures of self-efficacy (Marsh et al., 1991). However, measures included in the large cross-national OECD database used in Study 1 were not based on responses in which the response scale was clearly “absolute” as in traditional self-efficacy scales. In order to provide a more direct test of these predictions, Study 2 evaluated the BFLPE for math agency and MTSP. Although agency is sometimes considered to be a central component of self-efficacy, the nature of the items (see Appendix A) clearly invokes the use of comparative information and, at least in this respect, is more similar to typical operationalizations of self-concept than to measures of test-specific self-efficacy. Hence, it is not surprising that the results based on math agency provided clear support for the BFLPE (positive effects of individual achievement and negative effects of class-average achievement). More importantly in relation to our claims about the theoretical importance of the distinction between absolute and normative comparisons, the effect of class-average achievement was not statistically significant for MTSP responses—there was no BFLPE. In line with the results of Study 1, the results of Study 2 provide strong support for our claim that the size of the BFLPE depends on the nature of the specific items used to measure different self-belief constructs. Furthermore, the extent to which items invoke social comparison processes is a potentially important source of variation in the way that researchers construct other psychosocial constructs that has not been given sufficient attention in previous research.

### *Choice of Comparison Targets and the BFLPE*

Our research is apparently the first to demonstrate the simultaneous, independent negative predictive effects of two sources of social comparison information; the negative predictive effect of class- or school-average ability (the BFLPE) widely reported in educational psychology research and negative predictive effects of upward comparison strategies with individual target persons widely studied in SCT studies in social psychology research. In this respect, we bring together two distinct areas of research typically based on very different experimental designs and methodology, showing that each has the potential to contribute to the other.

An important focus of SCT research in social psychology has been the strategies that individuals use to select comparison target persons (e.g., upward and downward comparison strategies) to maximize competing needs. In BFLPE research, this target comparison is assumed to be a generalized other, as operationalized by class- or school-average achievement. Whereas this may be a reasonable assumption within the imposed social comparison paradigm (the basis of BFLPE research), even in this highly constrained environment, individual students have considerable flexibility in choosing individual classmates who might serve as a basis for social comparison. In this respect, the results of Study 2 are important as they demonstrate that both sources of social comparison information (class-average responses and upward comparisons) contribute negatively to self-perceptions of math agency.

It is important to note that each of these sources of social comparison information provides a unique, independent predictive effect that cannot be explained by the other. In this sense, the comparison target classmate selected by a student is more than just a “noisy” reflection of the class average as a basis of comparison (i.e., a “class average” based on a response of one randomly selected student that would obviously have considerable random error compared with the class average based on all students within the class). Had the target comparison classmate merely been a randomly selected student, then the effect of upward comparison in Study 2 would have not made an independent contribution after controlling for the effect of class-average achievement. Although consistent with a priori predictions based on the BFLPE, the results have important implications for both BFLPE and SCT research. Importantly, the uses of generalized and specific others are not mutually exclusive alternatives. Individuals might simultaneously evaluate their performances in relation to both the performances of specific target individuals selected in ways that have been considered in SCT research and in relation to some generalized other performance based on an average performance, as posited in BFLPE research. Hence, there is a need for more research to juxtapose different operationalizations used in SCT and BFLPE research. In particular, BFLPE studies should evaluate microlevel social comparison strategies used by individual students in their selection of target comparison classmates, whereas SCT research should incorporate macrolevel social comparison strategies based on class-average information (or some alternative representation of the “generalized other”) as well as microlevel strategies that have been the focus of this research. In relation to both operationalizations, there seems to be an important role for mixed-methods research in which the largely quantitative approach used in this research is supplemented with qualitative research to more fully explicate these alternative social comparison processes.

### *Limitations and Directions For Further Research*

As with all research studies, it is relevant to note potentially important limitations that may influence our interpretations of the results as well as providing useful directions for future research. Particularly in large-scale studies like those considered here, there is often an emphasis on using brief measures. This is clearly the case with the OECD PISA Student Approaches to Learning instrument (Marsh, Hau, et al., 2006) as the constructs were measured using brief scales, typically items selected from longer scales. This was particularly a concern with the MTSP construct in Study 2 that was based on a single response (a postdiction estimate of the number of test items a student had answered correctly). Although somewhat cumbersome in a survey study, it might be possible for students to estimate the probability that they answered individual items correctly as a basis of providing multiple indicators—a strategy like that used in many self-efficacy measures in educational settings (e.g., Marsh et al., 1991).

Previous BFLPE research, as was the case in the present investigation, has identified the BFLPE in relation to both school- and class-average ability. The theoretical model posits the contextual effects based on group-average achievement so that predictions are similar whether the “group” refers to the school or the class. Because class- and school-average ability are usually highly related (and confounded), it is typically difficult to separate out the



effects of one from the other. However, if there is clearly delineated ability grouping within schools such that there are large, systematic differences in class-average ability within the same school, it is reasonable to expect class-average ability effects in addition to school-average effects. Nevertheless, BFLPE studies have typically been unable to clearly distinguish between the effects of class-level achievement and school-level achievement (see discussion by Trautwein et al., 2006). Because there was no class-level information in Study 1 or school-level information in Study 2, we were not able to pursue this distinction in the present investigation. It is, however, a relevant issue for consideration in further research.

The results of the present investigation suggest that the BFLPEs were reasonably consistent across countries in the large cross-national PISA study (based on the very small country-to-country variation shown in Table 1; see further discussion by Marsh & Hau, 2003). Whereas the sample of countries is diverse, we note that Western countries are overrepresented. This is particularly relevant in that the current round of PISA data (collected in 2006 and released in late 2007) is even more diverse, with a larger number of countries overall and more non-Western countries. Hence, this new database should provide an even stronger basis for evaluating the cross-national generalizability of the BFLPE and country-level variables that might explain any such variation found between countries. We also note that most BFLPE research—like the present investigation—is based on responses by adolescents in high school settings (but see Craven, Marsh, & Print, 2000; Marsh, Chessor, Craven, & Roche, 1995). Particularly in relation to new issues raised here—the distinction between normative and absolute frames of reference and the distinction between individual- and group-average (generalized other) bases of social comparison—it is relevant to explore developmental differences and the role of social comparison in the formation of self-concept by young children.

An important contribution of the present investigation is the distinction between normative and absolute frames of reference and the implications for social comparison processes. Although our results are clearly consistent with this distinction, systematically manipulating the wording of the items and the response language for the same construct would provide a stronger test. Thus, for example, self-efficacy items could be expressed in terms of absolute standards (e.g., probability of getting an item correct, the percentage of test items that will be correctly answered, the amount of time taken to run 100 m), relative standards (how well one will do on a test relative to other students, what place one will get in a 100-m race), or ambiguous frames of reference (doing well on a test, doing well on a 100-m race). In this way it would be possible to unconfound the content of the construct from the nature of the items and the responses used to assess it. This is a particularly relevant consideration in self-efficacy research in which the nature of the self-efficacy construct seems to imply an absolute frame of reference but actual measurement instruments are sometimes based on a normative frame of references.

An important contribution of our research was to demonstrate that upward comparisons based on the target comparison classmate made an independent contribution to the prediction of ASC beyond what could be explained in terms of negative effects of class-average ability (the BFLPE). However, an important direction of future research is to refine this measure of target comparison

classmate and more fully explicate the psychological processes involved. Thus, for example, it might be useful to compare these results with those based on randomly assigned target comparison classmates who are not actually chosen by the student, thereby unconfounding the manner and direction of choice. It would also be useful to use a mixed-methods design that combines the quantitative approach emphasized here with a more qualitative approach to better understand the processes used to make this choice. Furthermore, whilst SCT research sometimes defines direction of comparison in terms of actual performance differences (based on the test completed as part of the study or previous achievement measures), we emphasized perceived differences as being more critical in terms of self-concept formation. Although perceived and actual differences are likely to be highly correlated, a useful direction of future research would be to more fully explore this distinction.

### *Multilevel Modeling Perspective*

BFLPE research has increasingly used a multilevel modeling approach in which it is possible to appropriately consider variables measured at the individual and group level, and to distinguish between effects that occur at different levels. This is a natural corollary of a simultaneous focus on the differential effects of individual achievement and class- or school-average achievement. A multilevel perspective is not so obviously relevant to SCT research that incorporates measures only at the individual level. However, SCT may also involve multilevel data in which individuals are nested within groups that occur naturally (e.g., neighborhoods, schools, institutions) or are constructed by the experimenter as part of the study. Not only is multilevel modeling a more appropriate way to analyze such multilevel data, but it also opens up new research questions not obvious to researchers who ignore the multilevel nature of their data (e.g., the extent to which the results generalize over different groups and group-level variables such as climate that might be related to individual-level variables or account for group-to-group differences in the effects of social comparison information). More generally, a multilevel perspective would facilitate the juxtaposition of individual- and group-level bases of comparison that have been emphasized in SCT and BFLPE research, respectively.

### References

- Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- Aspinwall, L. G., & Taylor, S. E. (1993). Effects of social comparison direction, threat, and self-esteem on affect, self-evaluation, and expected success. *Journal of Personality and Social Psychology*, 64, 708–722.
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.
- Bandura, A. (2006). Toward a psychology of human agency. *Perspectives on Psychological Science*, 1, 164–180.
- Baumert, J., Roeder, P. M., Sang, F., & Schmitz, B. (1986). Leistungsentwicklung und Ausgleich von Leistungsunterschieden in Gymnasialklassen [Achievement trajectories and reduction of achievement differences]. *Zeitschrift für Pädagogik*, 32, 639–660.
- Blanton, H., Buunk, B. P., Gibbons, F. X., & Kuyper, H. (1999). When better-than-others compare upward: Choice of comparison and comparative evaluation as independent predictors of academic performance. *Journal of Personality and Social Psychology*, 76, 420–430.



- Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review*, 15, 1–40.
- Branden, N. (1994). *The six pillars of self-esteem*. New York: Bantam Books.
- Bruner, J. (1996). A narrative model of self construction. *Psyke & Logos*, 17, 154–170.
- Buckingham, J. T., & Alicke, M. D. (2002). The influence of individual versus aggregate social comparison and the presence of others on self-evaluations. *Journal of Personality and Social Psychology*, 83, 1117–1130.
- Buunk, B. P., & Gibbons, F. X. (2000). Toward an enlightenment in social comparison theory: Moving beyond classic and Renaissance approaches. In J. Suls & L. Wheeler (Eds.), *Handbook of social comparison: Theory and research* (pp. 487–500). Dordrecht, Netherlands: Kluwer Academic.
- Craven, R. G., Marsh, H. W., & Print, M. (2000). Selective, streamed, and mixed-ability programs for gifted students: Impact on self-concept, motivation, and achievement. *Australian Journal of Education*, 44, 51–75.
- Diener, E., & Fujita, F. (1997). Social comparison and subjective well-being. In B. P. Buunk & F. X. Gibbons (Eds.), *Health, coping, and well-being: Perspectives from social comparison theory* (pp. 329–358). Mahwah, NJ: Erlbaum.
- Enders, C. K. (2006). Analyzing structural equation models with missing data. In G. R. Hancock & R. O. Mueller (Eds.), *Introduction to structural equation modeling: A second course* (pp. 313–344). Mahwah, NJ: Erlbaum.
- Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7, 117–140.
- Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Hodder Arnold.
- Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology*, 92, 160–170.
- Huguet, P., Dumas, F., Monteil, J. M., & Genestoux, N. (2001). Social comparison choices in the classroom: Further evidence for students' upward comparison tendency and its beneficial impact on performance. *European Journal of Social Psychology*, 31, 557–578.
- Husén, T. (1967). *International study of achievement in mathematics: A comparison of 12 countries* (Vols. 1–2). Stockholm, Sweden: Almqvist & Wiksell.
- Isaacson, R. M., & Fujita, M. (2006). Metacognitive knowledge monitoring and self-regulated learning: Academic success and reflections on learning. *Journal of the Scholarship of Teaching and Learning*, 6, 39–55.
- Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
- Little, T. D., Oettingen, G., Stetsenko, A., & Baltes, P. B. (1995). Children's action control beliefs about school performance: How do American children compare with German and Russian children? *Journal of Personality and Social Psychology*, 69, 686–700.
- Lüdtke, O., Köller, O., Marsh, H. W., & Trautwein, U. (2005). Teacher frame of reference and the big-fish-little-pond effect. *Contemporary Educational Psychology*, 30, 263–285.
- Major, B., Testa, M., & Bylsma, W. H. (1991). Responses to upward and downward social comparisons: The impact of esteem-relevance and perceived control. In J. Suls & T. A. Wills (Eds.), *Social comparison: Contemporary theory and research* (pp. 237–260). Hillsdale, NJ: Erlbaum.
- Marsh, H. W. (1990). *Self Description Questionnaire (SDQ) II: A theoretical and empirical basis for the measurement of multiple dimensions of adolescent self-concept: An interim test manual and a research monograph*. San Antonio, TX: Psychological Corporation.
- Marsh, H. W. (1991). The failure of high-ability high schools to deliver academic benefits: The importance of academic self-concept and educational aspirations. *American Educational Research Journal*, 28, 445–480.
- Marsh, H. W. (1992). *Self Description Questionnaire (SDQ) II: A theoretical and empirical basis for the measurement of multiple dimensions of adolescent self-concept: An interim test manual and a research monograph*. Campbelltown, New South Wales, Australia: SELF Research Centre, University of Western Sydney.
- Marsh, H. W. (1993). Academic self-concept: Theory, measurement, and research. In J. Suls (Ed.), *Psychological perspectives on the self* (Vol. 4, pp. 59–98). Hillsdale, NJ: Erlbaum.
- Marsh, H. W. (2007). *Self-concept theory, measurement and research into practice: The role of self-concept in educational psychology*. Leicester, England: British Psychological Society.
- Marsh, H. W., Chessor, D., Craven, R. G., & Roche, L. (1995). The effects of gifted and talented programs on academic self-concept: The big fish strikes again. *American Educational Research Journal*, 32, 285–319.
- Marsh, H. W., & Craven, R. G. (2002). The pivotal role of frames of reference in academic self-concept: The "big-fish-little-pond" effect. In F. Pajares & T. Urdan (Eds.), *Adolescence and education: Vol. 2. Academic motivation of adolescents*. Greenwich, CT: Information Age Publishing.
- Marsh, H. W., & Craven, R. G. (2006). Reciprocal effects of self-concept and performance from a multidimensional perspective: Beyond seductive pleasure and unidimensional perspectives. *Perspectives on Psychological Science*, 1, 133–163.
- Marsh, H. W., & Hau, K. T. (2003). Big-fish-little-pond effect on academic self-concept: A cross-cultural (26-country) test of the negative effects of academically selective schools. *American Psychologist*, 58, 364–376.
- Marsh, H. W., Hau, K. T., Artelt, C., Baumert, J., & Peschar, J. L. (2006). OECD's brief self-report measure of educational psychology's most useful affective constructs: Cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing*, 6, 311–360.
- Marsh, H. W., Kong, C-K., & Hau, K. T. (2000). Longitudinal multilevel modeling of the big-fish-little-pond effect on academic self-concept: Counterbalancing social comparison and reflected glory effects in Hong Kong high schools. *Journal of Personality and Social Psychology*, 78, 337–349.
- Marsh, H. W., & Rowe, K. J. (1996). The negative effects of school-average ability on academic self-concept—an application of multilevel modeling. *Australian Journal of Education*, 40, 65–87.
- Marsh, H. W., Tracey, D. K., & Craven, R. G. (2006). Multidimensional self-concept structure for preadolescents with mild intellectual disabilities: A hybrid multigroup-mimic approach to factorial invariance and latent mean differences. *Educational and Psychological Measurement*, 66, 795–818.
- Marsh, H. W., Walker, R., & Debus, R. (1991). Subject-specific components of academic self-concept and self-efficacy. *Contemporary Educational Psychology*, 16, 331–345.
- Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–26). Cambridge, MA: MIT Press.
- Organisation for Economic Cooperation and Development. (2001a). *Knowledge and skills for life: Results from the first OECD programme for international student assessment (PISA) 2000*. Paris: Author.
- Organisation for Economic Cooperation and Development. (2001b). *PISA international database 2000* [Data file]. Paris: Author.
- Organisation for Economic Cooperation and Development. (2003). *Student engagement at school: A sense of belonging and participation*. Paris: Author.
- Pajares, F., & Schunk, D. H. (2005). Self-efficacy and self-concept beliefs: Jointly contributing to the quality of human life. In H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.), *International advances in self*



- research (Vol. 2, pp. 95–123). Greenwich, CT: Information Age Publishing.
- Rasbash, J., Steele, F., Browne, W., & Prosser, B. (2004). *A user's guide to MLwiN - Version 2.0* [Software manual]. London: Institute of Education, University of London.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.) [Software manual]. Thousand Oaks, CA: Sage.
- Reuman, D. A. (1989). How social comparison mediates the relation between ability-grouping practices and students' achievement expectancies in mathematics. *Journal of Educational Psychology*, 81, 178–189.
- Robitaille, D., & Garden, R. (1989). *The IEA study of mathematics II: Contents and outcomes of school mathematics*. Oxford, England: Pergamon Press.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177.
- Seaton, M., Marsh, H. W., Dumas, F., Huguet, P., Monteil, J., Régner, I., et al. (in press). In search of the big fish: A reanalysis investigating the coexistence of the big-fish-little-pond effect with the positive effects of upward comparisons. *British Journal of Social Psychology*.
- Skaalvik, E. M., & Bong, M. (2003). Self-concept and self-efficacy revisited: A few notable differences and important similarities. In H. W. Marsh, R. G. Craven, & D. M. McInerney (Eds.), *International advances in self research* (Vol. 1, pp. 67–90). Greenwich, CT: Information Age Publishing.
- Skaalvik, E. M., & Rankin, R. J. (1995). A test of the internal/external frame of reference model at different levels of math verbal self-perception. *American Educational Research Journal*, 32, 161–184.
- Skaalvik, E. M., & Skaalvik, S. (2002). Internal and external frames of reference for academic self-concept. *Educational Psychologist*, 37, 233–244.
- Skinner, E. A., Zimmer-Gembeck, M. J., & Connell, J. P. (1998). Individual differences and the development of perceived control. *Monographs of the Society for Research in Child Development*, 63 (2–3, Serial No. 254).
- Smith, W. P., & Arnelsson, G. B. (2000). Stability of related attributes and the inference of ability through social comparison. In J. Suls & L. Wheeler (Eds.), *Handbook of social comparison: Theory and research* (pp. 45–66). Dordrecht, Netherlands: Kluwer Academic.
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.
- Stapel, D. A., & Suls, J. (2004). Method matters: Effects of explicit versus implicit social comparisons on activation, behavior, and self-views. *Journal of Personality and Social Psychology*, 87, 860–875.
- Suls, J. M., & Tesch, F. E. (1978). Students' preferences for information about their test performance: A social comparison study. *Journal of Applied Social Psychology*, 8, 189–197.
- Suls, J. M., & Wheeler, L. (2000). A selective history of classic and neo-social comparison theory. In J. M. Suls & L. Wheeler (Eds.), *Handbook of social comparison: Theory and research* (pp. 3–22). Dordrecht, Netherlands: Kluwer Academic.
- Swallow, S. R., & Kuiper, N. A. (1992). Mild depression and frequency of social comparison behavior. *Journal of Social & Clinical Psychology*, 11, 167–180.
- Tobias, S., & Everson, H. (2002). *Knowing what you know and what you don't: Further research on metacognitive knowledge monitoring* (Report No. 2002–3). New York: College Board.
- Trautwein, U., & Lüdtke, O. (2005). The Big Fish Little Pond Effect: Future research questions and educational implications. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology*, 19, 137–140.
- Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology*, 98, 788–806.
- Wheeler, L. (1966). Motivation as a determinant of upward comparison. *Journal of Experimental Social Psychology*, 2(Suppl. 1), 27–31.
- Wood, J. V. (1989). Theory and research concerning social comparisons of personal attributes. *Psychological Bulletin*, 106, 231–248.
- Wood, J. V. (1996). What is social comparison and how should we study it? *Personality and Social Psychology Bulletin*, 22, 520–537.
- Zimmerman, B. J. (1998). Developing self-fulfilling cycles of academic regulation: An analysis of exemplary instructional models. In D. Schunk & B. Zimmerman (Eds.), *Self-regulated learning: From teaching to self-reflective practice* (pp. 1–19). New York: Guilford Press.

## Appendix A

## Summary of Scales Used in Studies 1 and 2

## Scale

## Study 1

## Control Strategies

1. When I study, I start by figuring out exactly what I need to learn.
2. When I study, I force myself to check to see if I remember what I have learned.
3. When I study, I try to figure out which concepts I still haven't really understood.
4. When I study, I make sure that I remember the most important things.
5. When I study, and I don't understand something I look for additional information to clarify this.

## Effort and Perseverance

6. When studying, I work as hard as possible.
7. When studying, I keep working even if the material is difficult.
8. When studying, I try to do my best to acquire the knowledge and skills taught.
9. When studying, I put forth my best effort.

## Perceived Self-Efficacy

10. I'm certain I can understand the most difficult material presented in texts.
11. I'm confident I can understand the most complex material presented by the teacher.
12. I'm confident I can do an excellent job on assignments and tests.
13. I'm certain I can master the skills being taught.

## Control Expectation

14. When I sit myself down to learn something really difficult, I can learn it.
15. If I decide not to get any bad grades, I can really do it.
16. If I decide not to get any problems wrong, I can really do it.
17. If I want to learn something well, I can.

## Academic Self-Concept

18. I learn things quickly in most school subjects.
19. I'm good at most school subjects.
20. I do well in tests in most school subjects.

## Study 2

## Math Agency

1. When it comes to math, I'm pretty smart.
2. I can learn the things I need for math pretty fast, without really trying a lot.
3. I'm pretty smart in mathematics—even without working very hard.

(Appendixes continue)



Appendix B

Confirmatory Factor Analysis of Self-Belief Constructs in Study 1

Item	CStrg	Effr	SEff	CExp	ASC	Unique
Factor loadings						
1	.524					.726
2	.634					.598
3	.681					.536
4	.671					.550
5	.596					.645
6		.655				.572
7		.677				.542
8		.730				.467
9		.674				.546
10			.613			.624
11			.687			.528
12			.684			.532
13			.730			.467
14				.622		.613
15				.660		.564
16				.621		.615
17				.719		.483
18					.659	.566
19					.782	.388
20					.781	.389
Factor correlations <sup>a</sup>						
CStrg	1	.703	.545	.543	.400	
Effr	.926	1	.549	.547	.433	
SEff	.728	.741	1	.666	.552	
CExp	.737	.725	.903	1	.490	
ASC	.518	.551	.705	.637	1	

*Note.* CStrg = Control Strategies; Effr = Effort and Persistence; SEff = Self-efficacy; CExp = Control Expectations; ASC = Academic Self-Concept, Unique = uniqueness (error). See Appendix A for the wording of the items. Goodness of fit for this confirmatory factor analysis was very good: root mean square error of approximation (RMSEA) = 0.055; non-normed fit index (NNFI) = 0.978; comparative fit index (CFI) = 0.981; standardized root mean square residual (SRMR) = 0.035. See Marsh, Hau, Artlet, et al. (2006) for more detail on the factor structure and results from multiple group confirmatory factor analyses showing that the factor structure is invariant across countries participating in the Organisation for Economic Cooperation and Development Program of Student Assessment data collection.

<sup>a</sup> Correlations below the main diagonal are based on this confirmatory factor analysis, whereas the correlations above the main diagonal are scale correlations based on the actual scale scores used in the study.

Received December 12, 2006  
Revision received September 18, 2007  
Accepted October 31, 2007 ■

# Longitudinal Analysis of the Role of Perceived Self-Efficacy for Self-Regulated Learning in Academic Continuance and Achievement

Gian Vittorio Caprara, Roberta Fida,  
Michele Vecchione, Giannetta Del Bove,  
Giovanni Maria Vecchio, and Claudio Barbaranelli  
“Sapienza” University of Rome

Albert Bandura  
Stanford University

The present study examined the developmental course of perceived efficacy for self-regulated learning and its contribution to academic achievement and likelihood of remaining in school in a sample of 412 Italian students (48% males and 52% females ranging in age from 12 to 22 years). Latent growth curve analysis revealed a progressive decline in self-regulatory efficacy from junior to senior high school, with males experiencing the greater reduction. The lower the decline in self-regulatory efficacy, the higher the high school grades and the greater the likelihood of remaining in high school controlling for socioeconomic status. Reciprocal cross-lagged models revealed that high perceived efficacy for self-regulated learning in junior high school contributed to junior high school grades and self-regulatory efficacy in high school, which partially mediated the relation of junior high grades on high school grades and the likelihood of remaining in school. Socioeconomic status contributed to high school grades only mediational through junior high grades and to school drop out both directly and mediational through junior high grades.

**Keywords:** perceived self-regulatory efficacy, self-regulated learning, academic achievement, latent growth curves, school drop out

The present study analyzed the developmental course of perceived self-efficacy for self-regulated learning from junior high to high school and its contribution to academic achievements and the likelihood of remaining in school. The role of perceived self-efficacy in the academic domain has been examined at three different levels. These include students' beliefs in their efficacy to regulate their learning activities and to master academic subjects, teachers' beliefs in their instructional efficacy to manage classrooms and to motivate and promote learning in their students, and faculties' collective sense of efficacy that their schools can accomplish significant academic progress. These different lines of theorizing and research have been reviewed in some detail (Bandura, 1997; Pajares & Urdan, 2006; Schunk & Pajares, 2002).

The present longitudinal study focuses on the central role played by perceived self-regulatory efficacy in one's academic self-development and functioning. The capacity to regulate one's thoughts, motivation, affect, and action through self-reactive in-

fluence constitutes one of the core properties of human agency within the conceptual framework of social cognitive theory (Bandura, 2006b; in press). For linguistic brevity, perceived self-efficacy for self-regulated learning will be referred to as *self-regulatory efficacy*.

Self-regulatory efficacy was selected as a key factor because of its growing primacy in contemporary life. Information technologies are globalizing knowledge and altering educational systems (Bandura, 2002). In the past, students' educational development depended on the quality of the schools in which they were enrolled. Students can now exercise greater personal control over their own learning, independently of time and place, through multimedia instruction on the Internet. In this new era, the construction of knowledge will rely increasingly on electronic inquiry. In research in self-instruction through the Internet, students with high efficacy for self-regulated learning are the ones who make the best use of Internet-based instruction (Debowski, Wood, & Bandura, 2001; Joo, Bong, & Choi, 2000). Moreover, the accelerated pace of social, informational, and technological change is placing a premium on capability for self-directed learning and self-renewal. People now have to educate themselves throughout their lifetime.

Within the agentic framework of social cognitive theory, self-regulation operates through three generic subfunctioning (Bandura, 1986; in press). These include self-monitoring of one's activities and the cognitive and social conditions under which one engages in them; adoption of proximal goals, rooted in a value system, to motivate and guide one's efforts and the strategies for realizing the challenges set for oneself; and the exercise of self-influence that includes the enlistment of self-motivating incentives and social supports to sustain one's academic pursuits. Different

---

Gian Vittorio Caprara, Roberta Fida, Michele Vecchione, Giannetta Del Bove, Giovanni Maria Vecchio, and Claudio Barbaranelli, Department of Psychology, “Sapienza” University of Rome, Rome, Italy; Albert Bandura, Department of Psychology, Stanford University.

This study was partially supported by grants from the Ministry of Education, University and Research (PRIN, 2002/2004); University of Rome La Sapienza, Ateneo Research in 2002, 2003, and 2004 to Gian Vittorio Caprara; and from the Spencer Foundation and W. T. Grant Foundation to Albert Bandura.

Correspondence concerning this article should be addressed to Gian Vittorio Caprara, Department of Psychology, “Sapienza” University of Rome, Via dei Marsi 78, 00185, Rome, Italy. E-mail: Gianvittorio.Caprara@uniroma1.it



models of self-regulation have been proposed (Schunk & Zimmerman, 1994; Zimmerman & Schunk, 1989). Although the models differ in particulars, they generally include self-assessment through self-monitoring, instrumental cognitive and metacognitive guides, goal setting, and self-motivational strategies.

Analyses of the role of self-regulation in the acquisition of knowledge and cognitive skills have been largely confined to enhancement of academic learning by use of task-related metacognitive strategies. A number of theorists have addressed the pragmatics of self-regulation in terms of selecting appropriate strategies, testing one's comprehension and state of knowledge, correcting one's deficiencies, and recognizing the utility of cognitive strategies (Brown, 1987; Paris & Newman, 1990). Self-directive use of cognitive strategies is a part of the way in which students regulate their own cognitive development and functioning. Social cognitive theory integrates the cognitive and metacognitive factors with motivational self-regulation mechanisms (Bandura, 1986; Zimmerman, 2000; Zimmerman & Cleary, 2006). This theory expands the conception of self-regulation in two directions. First, it incorporates a larger set of self-regulatory mechanisms governing cognitive functioning. Second, it encompasses social and motivational skills as well as cognitive ones.

Zimmerman (1989, 2000) has been the leading exponent of an expanded model of academic self-regulation. Viewed within the conceptual framework of social cognitive theory, people must develop skills to regulate the motivational, affective, and social determinants of their intellectual functioning as well as the cognitive aspects. This requires bringing self-influence to bear on every aspect of their learning experiences. There is a major difference between possessing self-regulatory knowledge and skills and being able to put them into practice and to stick with them. Self-regulatory skills will not contribute much if students cannot get themselves to apply them persistently in the face of difficulties, stressors, and competing attractions. Firm belief in one's self-regulatory efficacy provides the staying power. Children's belief that they can regulate their own learning raises their efficacy for academic activities. Their academic efficacy increases their achievement both directly and by raising their academic aspirations (Zimmerman & Bandura, 1994; Zimmerman, Bandura, & Martinez-Pons, 1992).

The belief that people hold about their capabilities affects the quality of their functioning through four major processes: cognitive, motivational, affective, and decisional (Bandura, 1997). The independent contribution of efficacy beliefs to cognitive functioning is verified experimentally by Bouffard-Bouchard (1990) in research in which high or low self-efficacy beliefs were instilled arbitrarily in students irrespective of their actual performance. Students whose sense of efficacy was raised set higher aspirations for themselves, showed greater strategic flexibility in the search for solutions, achieved higher intellectual performances, and were more accurate in evaluating the quality of their performances than were students of equal cognitive ability who were led to believe they lacked such capabilities. Efficacy beliefs contributed to accomplishments both motivationally and through support of strategic thinking. Self-regulatory efficacy also raises academic goals and aspirations, personal standards for the quality of work considered to be acceptable, and beliefs in one's capabilities for academic achievement after one controls for instructional level, prior aca-

demical performance, and relevant aptitude (Zimmerman & Bandura, 1994; Zimmerman et al., 1992).

The present study focused on adolescence because it is an especially taxing transitional phase that presents a host of new challenges (Bandura, 2006a; Graber, Brooks-Gunn, & Petersen, 1996; Pajares & Urdan, 2006). Adolescents have to manage major biological, educational, and social role transitions concurrently. Learning how to deal with pubertal changes, differently structured school environments, enlarged peer networks, and emotionally invested partnerships becomes important. Moreover, this is the time when the roles of adulthood must begin to be addressed in almost every dimension of life. Adolescents must also begin to consider seriously what they want to do with their lives (Bandura, Barbaranelli, Caprara, & Pastorelli, 2001). They have to master many new skills and the ways of adult society. The way in which adolescents develop and exercise their personal efficacy during this period can play a key role in setting the course their life paths take (Bandura, 2006b; Pajares & Urdan, 2006).

The transition from middle-level school to high school involves a major environmental change that can tax personal efficacy (Wigfield, Eccles, Schiefele, Roeser, & Davis-Kean, 2006). Under the new social structural arrangements, the adolescents have to reestablish their sense of efficacy, social connectedness, and status within an enlarged heterogeneous network of new peers. During this adaptational period, young adolescents sense some loss of personal control, become less confident in themselves, and suffer some decline in self-motivation (Eccles & Midgley, 1989). But these adverse effects are neither universal nor enduring for every adolescent. It was predicted that adolescents who have a high sense of efficacy to regulate their learning activities in junior high school are likely to do better academically in high school and less likely to drop out of school.

We used a latent growth curve approach (Duncan & Duncan, 1995; McArdle, 1988; McArdle & Anderson, 1989; Meredith & Tisak, 1990; Stoolmiller, 1994) to investigate the level and stability of self-regulatory efficacy and the extent to which these beliefs at age 12 and their subsequent change can predict academic achievement at the end of high school in males and females.

For reasons given earlier, we predicted decline in self-regulatory efficacy after the transition from junior school to high school. In a cross-national study on a sample of 1,180 children ranging in age from 10 to 15 years (Pastorelli et al., 2001), girls exhibited a higher sense of efficacy to exercise control over their academic development than did boys. It was, therefore, expected that females would be more successful than males in maintaining their perceived efficacy to manage their academic activities as they progress in the educational system.

We used reciprocal cross-lagged relationships to evaluate the unique contribution of self-regulatory efficacy in junior high school to academic achievement in junior and senior high school and to continuing in school over and above the effects of prior academic performance and socioeconomic status (SES). Both experimental and naturalistic studies have shown that perceived academic self-efficacy makes an independent contribution after the effects of prior performance are partialled out (Bandura, 1997; Bouffard-Bouchard, 1990; Gore, 2006; Zimmerman & Martinez-Pons, 1986).

It was predicted that self-regulatory efficacy in junior high would contribute to academic achievement and continuance in



school through two pathways: by supporting a high sense of self-regulatory efficacy in high school and by mediating the effects of junior high academic performance and SES.

SES was selected because it can affect academic aspirations, availability of resources conducive to intellectual development, choice of peers who may support academic pursuits or disengagement from them, and the range of occupational pursuits that are seriously considered (Bandura, Barbaranelli, Caprara, & Pastorelli, 1996; Bandura et al., 2001; Sirin, 2005). Diverse lines of evidence show that SES affects performance, in large part, through its impact on psychosocial processes rather than directly (Bandura et al., 1996, 2001; Elder, 1995; Furstenberg, Cook, Eccles, Elder, & Sameroff, 1999). It was, therefore, predicted that SES influences academic achievement in high school and school drop out both directly and indirectly, through its impact on self-regulatory efficacy and academic performance in junior high school.

## Method

### Participants

The participants were part of an ongoing longitudinal project that began in 1989 designed to clarify the personal and social determinants of developmental pathways from childhood to early adulthood. A total of 412 children (196 males and 216 females) participated in the study. This longitudinal project used a staggered, multiple cohort design ranging from 1994 to 2004. The study includes two cohorts assessed at six different time points. Both cohorts were age 12 years at Time (T) 1, age 13 at T2, age 14 at T3, age 16 at T4, age 18 at T5, and age 20 and 22 at T6. Cohort effects were tested and found to be nonsignificant on all variables examined in this study. Therefore, the data from the two cohorts were combined.

Participants were originally drawn from the two public junior high schools in a community located near Rome. This sample represents a socioeconomic microcosm of the larger Italian society, composed of families of skilled workers, farmers, professionals, local merchants, and their service staff. In particular, 16% of families were in professional or managerial ranks, 37% were merchants or employees in various types of businesses, 14% were skilled workers, 31% were unskilled workers, 1% were retired, and 1% were unemployed. This occupational socioeconomic distribution matches the national profile (Istituto Italiano di Statistica, 2002). The composition of the family also matches national data with regard to type of families and number of children. Most children were from intact families (94%). The participation rate was high during the longitudinal data collection: 100% from T1 to T2 for both cohorts; 97% and 96% from T1 to T3 for first cohort and second cohort, respectively; 90% and 96% from T1 to T4 for first cohort and second cohort, respectively; 81% and 93% from T1 to T5 for first cohort and second cohort, respectively; 62% and 69% from T1 to T6 for first cohort and second cohort, respectively. Multivariate analysis of variance and Box's M test, carried out separately for males and females, revealed that there were no statistically significant differences on the means of the variables of interest (males:  $F[1, 143] = 1.34$ ,  $\Lambda = .95$ ,  $ns$ ; females:  $F[1, 165] = 1.69$ ,  $\Lambda = .95$ ,  $ns$ ) and on the covariance matrices (males:  $M = 32.24$ ,  $F[21, 38033.7] = 1.46$ ,  $ns$ ; females:  $M = 21.06$ ,  $F[21, 12531.5] = 0.93$ ,  $ns$ ) between the participants who provided com-

plete data for the present study and the ones who dropped out over the years.

### Procedures

At T1, T2, and T3 two experimenters administered in the classroom the scale measuring perceived efficacy for self-regulated learning. A stringent consent procedure for the research was followed including, at various stages, parents' consent and approval from school councils and freedom of children to decline participation if they choose to do so. The researchers explained that responses to the questionnaires would be confidential. At T4, T5, and T6 the participants received the questionnaire after being contacted by phone. Adolescents received a small payment (€25 or a dinner token) for their participation.

### Measures

*Perceived efficacy for self-regulated learning.* We assessed perceived efficacy for self-regulated learning (Bandura, 1990) from T1 to T5. This scale was previously validated on Italian samples (Bandura et al., 1996; Caprara, 2001) and American samples (Zimmerman et al., 1992). It includes 11 items that measure children's self-efficacy to plan and organize their academic activities (e.g., "How well can you organize your school work?"), to structure environments conducive to learning (e.g., "How well can you arrange a place to study without distractions?"), and to motivate themselves to do their school work (e.g., "How well can you study when there are other interesting things to do?"). For each item, participants rated the strength of their efficacy to execute the designed activities using a 5-point scale ranging from 1 (*cannot do at all*) to 5 (*highly certain can do*).

A principal-axis factor analysis revealed a one-factor structure at all time points. The percentage of the variance explained ranged from 33% to 38%. Cronbach's alpha was .83 at T1, .85 at T2, .86 at T3, .84 at T4, and .87 at T5.

*Academic achievement.* At the end of junior high school (T3), we assessed children's academic achievement for different subject matters (mathematics, science, language, and social studies) using a five-level grade system. We created a composite measure of academic achievement from the grades assigned by the group of teachers. In T6, high school academic achievement was assessed with a stringent examination system. The final high school grade was based on a national written exam supplemented with an oral exam. Students reported whether they graduated from high school and, if so, their final grade. In the Italian educational system, grades range from 60 to 100. Final high school grades and high school drop out (0 = drop out; 1 = graduated) served as the outcome variables.

As a check on students' reports of their high school grades, we compared the self-reported grades for a sample of 30 students against the grades recorded by the schools. In 29 of the 30 students, the self-reported grade was identical with the recorded school grades.

*SES.* Family SES was based on the occupation and education of the fathers and the mothers (see Sirin, 2005). We performed a confirmatory factor model, using the WLSMV as the method of



Table 1  
Means and Standard Deviations in Self-Regulatory Efficacy, Junior High School, and High School Grades

Self-regulatory efficacy	Males			Females		
	<i>n</i>	<i>M</i>	<i>SD</i>	<i>n</i>	<i>M</i>	<i>SD</i>
T1 (12 years old)	193	3.07	0.48	213	3.21	0.43
T2 (13 years old)	196	3.05	0.44	216	3.20	0.45
T3 (14 years old)	188	3.05	0.44	209	3.18	0.44
T4 (16 years old)	181	2.89	0.47	202	3.07	0.48
T5 (18 years old)	170	2.88	0.51	190	3.12	0.46
Junior high school grades (T4)	191	2.95	0.84	210	3.16	0.86
High school grades (T6)	77	75.34	13.07	132	80.13	12.63

Note. The items of the perceived efficacy for self-regulated learning scale were on a 5-point scale ranging from 1 (*cannot do at all*) to 5 (*highly certain can do*). Junior high school grades ranged from 1 to 5. High school grades ranged from 60 to 100.

estimation<sup>1</sup> (Muthén & Muthén, 1998), where SES was defined by parent’s education and occupation. After establishing the monodimensionality of this set of indicators (52% of variance explained;  $\alpha = .77$ ), we estimated the factor score of SES. This variable was included as observed time invariant covariate in the analysis.

Results

Descriptive Statistics

Observed means and standard deviations for self-regulatory efficacy across the five time points (from T1 to T5) are reported in Table 1 separately for males and females.

In longitudinal research it is common to have attrition over time (Hansen, Tobler, & Graham, 1990). With missing data, parameters’ estimation must be adjusted. Among the different methods used for taking into account missing data, we selected the most commonly used maximum-likelihood parameters’ estimation (Muthén & Shedden, 1999; Schafer & Graham, 2002). Thirty-three percent of males and 15% of females dropped out of the school.

Pairwise correlations (disattenuated for unreliability) between self-regulatory efficacy from T1 to T5 are provided in Table 2. They reveal a medium-to-high relative stability across time with lower relations the longer the elapsed time period.

Growth Models

The analysis of self-regulatory efficacy development was conducted within a latent variable framework. We specified a multi-group growth curve model that simultaneously estimated the same pattern of relationships among variables for males and females. To examine gender differences in the estimated parameters, we constrained all parameters to be equal across groups, and we used the chi-square difference test to compare nested models. Modification indices were used to assess the tenability of the equality constraint imposed across gender.

Two latent variables were specified from multiple indicators, that is, the five repeated measures of self-regulatory efficacy (from T1 to T5). The first factor is the intercept and it represents the baseline of self-regulatory efficacy (T1). The second factor is the slope or the shape of the trajectory over time and its mean gives the growth rate of self-regulatory efficacy.

The following equation shows the mathematical representation of the growth model:

$$y_t = \eta_0 + \eta_1 x_t + \varepsilon_t; \quad t = 1, 2, 3, 4, 5;$$

where  $y_t$  is the observed score at time  $t$ ,  $\eta_0$  is the unobserved score for the intercept factor,  $\eta_1$  is the unobserved score for the growth rate factor, and  $x_t$  is the factor loading relating  $y_t$  to latent growth variables.

Because factor loadings of the slope give the shape of the growth, alternative models were tested and compared with each other. We could establish the parameterization that provided the best fit to the data. We fixed the starting point for self-regulatory efficacy at T1 at 0 for all the models. Following McArdle and Anderson (1989), the first model tested was a no-growth model (with values for males followed by values for females),  $\chi^2(29, N = 196; 216) = 110.70, p < .001$ , comparative fit index (CFI) = .88, root-mean-square error of approximation (RMSEA) = .117 (.094; .140), standardized root-mean-square residual (SRMR) = .170. This model assumes that the level of self-regulatory efficacy is stable over time except for a random error component at each point of assessment. The second is a linear model representing a constant change over time,  $\chi^2(23, N = 196; 216) = 26.28, p = .29$ , CFI = 1.00, RMSEA = .026 (.000; .065), SRMR = .090. In this model we fixed the factor loadings at 0, 0.5, 1, 2, and 3. The third model examines a nonlinear growth where the form of the change over time is not specified a priori,  $\chi^2(20, N = 196; 216) = 19.70, p = .48$ , CFI = 1.00, RMSEA = .000 (.000; .059), SRMR = .083. Several parameters were added to the examined models. In particular, we specified the covariance between the residual variance of self-regulatory efficacy. Because these models are nested, we performed a chi-square differences test ( $\Delta\chi^2$ ) to compare the models (Bollen, 1989). This test revealed that the linear model provided the best fit to the data compared to the no-growth model,  $\Delta\chi^2(6) = 84.42, p < .001$ . The chi-square for the linear model was not statistically different from the chi-square for the nonlinear model,  $\Delta\chi^2(3) = 6.58, p = .086$ , but it was more parsimonious. Table 3 presents parameter estimates for the linear model.

The mean of the slope reveals a decline in self-regulatory efficacy over the time course for male and female subsamples.

<sup>1</sup> The WLSMV is a weighted least square parameter estimates that uses a diagonal weight matrix with robust standard errors and mean and variance adjusted chi-square test statistics (see Muthén & Muthén, 1998). This estimator is particularly suited for dealing with nonnormal or categorical data (Flora & Curran, 2004).

Table 2  
Correlations Among Self-Regulatory Efficacy Across Time

Variable	1	2	3	4	5
1. T1 (12 years old)	—	.71**	.70**	.57**	.52**
2. T2 (13 years old)	.67**	—	.76**	.64**	.59**
3. T3 (14 years old)	.63**	.59**	—	.68**	.68**
4. T4 (16 years old)	.50**	.47**	.60**	—	.81**
5. T5 (18 years old)	.27*	.43**	.50**	.58**	—

Note. Coefficients for females are above the diagonal; coefficients for males are below the diagonal. Correlations were disattenuated for unreliability.  
\*  $p < .05$ . \*\*  $p < .01$ .

However, the decrease was greater for males, as confirmed by a significant chi-square difference among constrained and unconstrained models,  $\Delta\chi^2(1) = 12.19$ ,  $p < .001$ . In particular self-regulatory efficacy decreases .077 for males and .035 for females each year. Moreover, the level of self-regulatory efficacy at baseline was higher for females,  $\Delta\chi^2(1) = 4.498$ ,  $p < .05$ . The variances of the growth factors were also estimated, and they indicated that there was a significant variation in individual differences both in the initial status and in the growth rate for males as well as for females. Figure 1 shows the trajectory for self-regulatory efficacy across junior high school and high school levels in the educational system for both males and females.

### Predicting High School Grades and School Drop Out

After establishing the best fitting growth curve model, we added SES as the time invariant covariate. To evaluate the contribution of the initial level of self-regulatory efficacy and its change over time on academic achievement, we tested two models. In the first model, we considered high school grades as the outcome. In the second model the probability of graduating from high school served as the second outcome.

The results of the first model are provided in Table 4. The posited model provided a good fit to the empirical data,  $\chi^2(41, N = 196; 216) = 52.90$ ,  $p = .10$ , CFI = .98, RMSEA = .038 (.000, .064), SRMR = .095. The higher the self-regulatory efficacy at T1, the higher the grades at the end of the high school. Moreover, the less self-regulatory efficacy declined from T1 to T5, the higher the

high school grades. The higher the SES the smaller the decline of self-regulatory efficacy. SES did not influence self-regulatory efficacy at T1. Although SES did not directly influence the grades, it contributed indirectly through its influence on self-regulatory efficacy change (total indirect effects were  $\beta = .11$ ,  $t = 3.93$  for males and  $\beta = .12$ ,  $t = 3.93$  for females). There were no significant differences on any of the parameters estimated in the samples of males and females. The model explains 17% of variance in high school grades for males and 21% of variance for females.

The results of the second model concerned with school drop out are provided in Table 5. The posited model also provided a good fit to the data,  $\chi^2(21, N = 107, 142) = 20.65$ ,  $p = .48$ , CFI = 1.00, RMSEA = .000; weighted root-mean-square residual = .89. The higher self-regulatory efficacy at T1, the lower the probability of dropping out of high school. Moreover, the more self-regulatory efficacy decline from T1 to T5, the higher the probability of dropping out of school. Furthermore, the higher the SES, the lower the probability of dropping out and the lower the decrease in self-regulatory efficacy. As in the case of grades, SES did not influence self-regulatory efficacy measured at T1. There were no significant differences on any the parameters estimated in the samples of males and females. The model explained 55% of variance of dropping out for males and 57% of variance for females.

### The Relation of Perceived Efficacy for Self-Regulated Learning to High School Grades and School Drop Out

We performed multigroup structural equation modeling to examine the role of self-regulatory efficacy, prior academic performance, and SES on high school grades and high school drop out. Except for a small variation in the role of SES in the two outcomes, the pattern of relations among the variables in the structural model are highly similar for high school grades and high school drop out. The structural relations among the variables are presented in Figures 2 and 3.

As shown in the figures, self-regulatory efficacy was relatively stable at the junior high level. Self-regulatory efficacy in junior high school contributed to both junior high grades and self-regulatory efficacy in high school. Self-regulatory efficacy in high school partially mediated the relation of junior high grades to high school grades (total indirect effects were  $\beta = .07$ ,  $t = 3.19$  for males and  $\beta = .07$ ,  $t = 3.19$  for females) and school drop out (total indirect effects were  $\beta = .06$ ,  $t = 2.13$  for males and  $\beta = .06$ ,  $t = 2.13$  for females). Self-regulatory efficacy thus contributed

Table 3  
Growth Curve Parameters for the Linear Model

Growth parameter	Males		Females	
	Parameter	<i>t</i>	Parameter	<i>t</i>
Mean				
Intercept	3.09	105.24	3.20	117.88
Slope	-.08	-5.56	-.03	-3.23
Variances				
Intercept	.11	5.98	.12	7.31
Slope	.01	3.59	.01	4.32
Correlation				
Intercept ↔ Slope	-.38	-2.23	-.26	-2.09

Note. The *t* values greater than 1.96 (1.65 for variances) indicate a parameter estimate that is significantly different from zero. Parameters estimated for correlations (↔) are presented in standardized form. All other parameter estimates are presented as unstandardized coefficients.



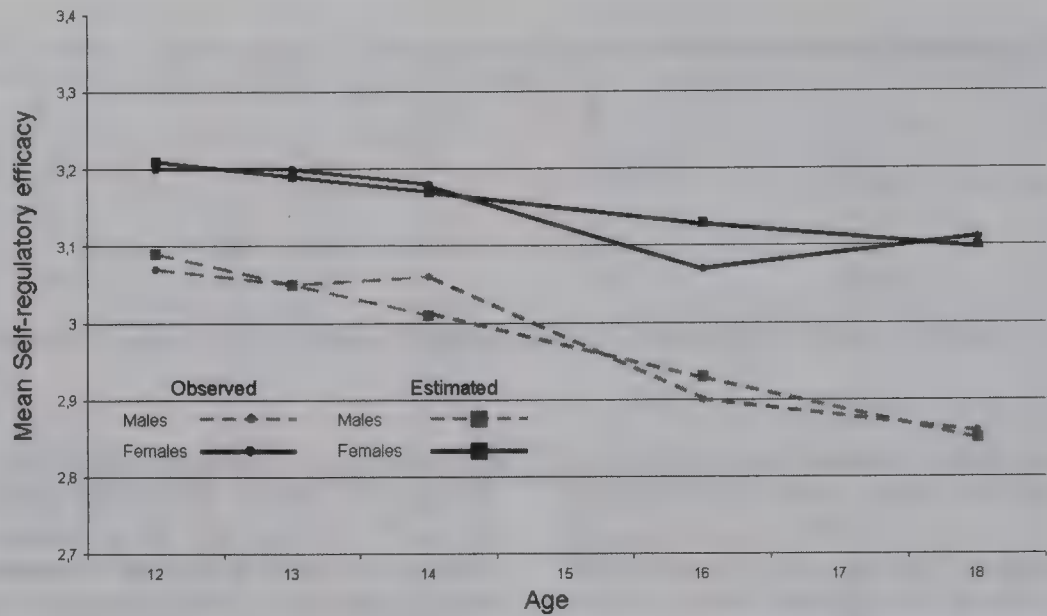


Figure 1. Differential change in self-regulatory efficacy in male and female students across different levels of education.

uniquely to high school grades and retention in school after we controlled for prior academic achievement.

Junior high grades were also related to high school grades and school drop out and completely mediated the relation of SES to high school grades (total indirect effects were  $\beta = .09$ ,  $t = 2.64$  for males and  $\beta = .10$ ,  $t = 2.64$  for females). In the case of school drop out, SES also contributed directly as well as through its relations to high school grades (total indirect effects were  $\beta = .13$ ,  $t = 4.13$  for males and  $\beta = .14$ ,  $t = 4.13$  for females).

The posited models provided a good fit to the empirical data. Regarding high school grades, the values for the various fit indices were  $\chi^2(17) = 14.85$ ,  $p = .61$ , CFI = 1.00, RMSEA = .000 (.000, .055), SRMR = .065. The model accounted for 25% of the variance in high school grades for males and 28% of the variance for females. The chi-square difference test revealed that all parameters were equal across gender, with the exception of the regression coefficient between self-regulatory efficacy during junior high school. This coefficient was larger for females indicating a slightly higher stability for females.

The model also provided a good fit to the empirical data for high school drop out,  $\chi^2(14) = 9.733$ ,  $p = .78$ , CFI = 1.00, RMSEA = .000, weighted root-mean-square residual = .720. The model

accounted for 53% of the variance of high school performance for males and 61% of the variance for females. The chi-square difference tests suggested that all parameters were equal across gender.

Discussion

Analysis of the trajectory of self-regulatory efficacy reveals a progressive decline as students advance through the educational system. Other studies have reported a similar decline but for students' beliefs in their efficacy for academic achievement, rather than self-regulatory efficacy (Britner & Pajares, 2006; Harter, 1996; Midgley, Feldlaufer, & Eccles, 1989). Several factors may account for students' loss of confidence in their capabilities to manage their academic activities. With increasing levels of schooling, the complexities of academic demands increase and cumulating scholastic deficits become increasingly salient. These changes confront students with adaptational pressures that inevitably shake their sense of efficacy. Students also gain new information about the nature of the academic activities, which provides them with a basis for reappraising their efficacy to get themselves to do them.

Table 4  
Predicting High School Grades: Structural Model

Regression	Males		Females	
	Parameter	t	Parameter	t
Intercept → Grades	.33	4.81	.37	4.81
Slope → Grades	.32	3.20	.34	3.20
SES → Intercept	.09	1.20	.09	1.20
SES → Slope	.18	2.20	.20	2.20
SES → Grades	.08	1.21	.09	1.21

Note. The  $t$  values greater than  $|1.96|$  indicate a parameter estimate that is significantly different from zero. Parameters estimated for regressions (→) are presented in standardized form. SES = socioeconomic status.

Table 5  
Predicting the Probability of Dropping Out of High School: Structural Model

Regression	Males		Females	
	Parameter	t	Parameter	t
Intercept → Drop out	.29	3.72	.30	3.72
Slope → Drop out	.46	2.94	.41	2.94
SES → Intercept	.06	0.76	.06	0.76
SES → Slope	.25	2.55	.29	2.55
SES → Drop out	.46	4.64	.49	4.64

Note. The  $t$  values greater than  $|1.96|$  indicate a parameter estimate that is significantly different from zero. Parameters estimated for regressions (→) are presented in standardized form. SES = socioeconomic status.

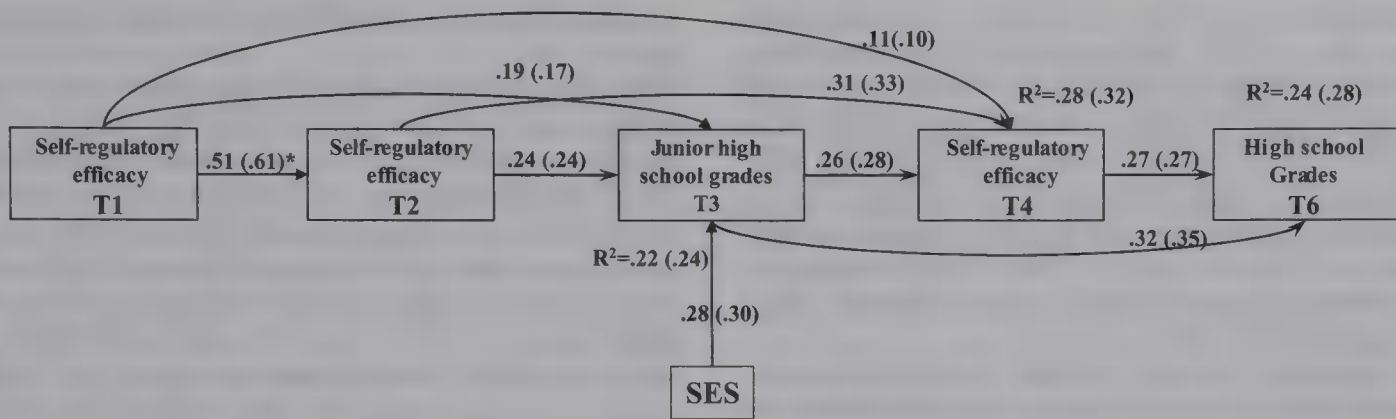


Figure 2. Contribution of self-regulatory efficacy to high school grades operating in conjunction with prior academic achievement and socioeconomic status (SES). The figure includes all of the coefficients that are significant beyond the  $p = .05$  level. The first coefficient in each structural link is for males, and the second coefficient in parentheses is for females.

Academic development is a product of a collaborative process within a social system rather than residing solely in students. Teachers also report a decline across grade level in their efficacy to motivate and promote their students' academic attainments (Bandura, 1997). Thus, students' adaptational problems are likely to be exacerbated if the teachers doubt they can achieve much success by their instructional efforts at higher grade levels (Midgley et al., 1989). Moreover, as students progress to late adolescence and young adulthood there are more competing activities that command their attention. Students report the lowest sense of efficacy to manage their academic activities when there are other interesting things to do (Zimmerman et al., 1992).

Both the initial level of self-regulatory efficacy and the degree of decline vary as a function of gender. Compared to male students, female students exhibit higher self-regulatory efficacy and a lesser decline as they progress in the educational system. The differential gender level is replicated cross-culturally. Female students in both Eastern and Western European countries exhibit higher perceived efficacy to regulate their academic activities than do male students (Pastorelli et al., 2001). The findings of the present study reveal that this gender gap widens as students progress in the educational system. Pastorelli and her collaborators

(2001) also found that students in more authoritarian systems perceive themselves as less efficacious for self-directed learning.

There are several possible explanations for the gender gap in self-regulatory efficacy. During the socialization process, girls are generally subjected to more social constraints on the range of activities they can engage in, especially outside the home, than are boys (Bussey & Bandura, 1999). Attractive competing options reduce opportunities to develop a sense of efficacy for self-directed academic learning and put a strain on efforts to stick to academic tasks. Differences in social and normative influences may also undermine the differential development and exercise of self-regulatory efficacy in the academic domain. Peer pressures for engagement in activities that compete with academic pursuits are likely to be stronger for boys than for girls (Jessor, Donovan, & Costa, 1991; Ogbu, 1990). Research is also needed to determine whether the educational system instills a belief of lesser academic self-regulatory efficacy in males, as suggested in research by Dweck and her collaborators (Dweck, Davidson, Nelson, & Enna, 1978).

The developmental trajectory is an aggregate measure for the students as a whole. It varies not only by gender but as a function of belief in one's efficacy to exercise some control over one's

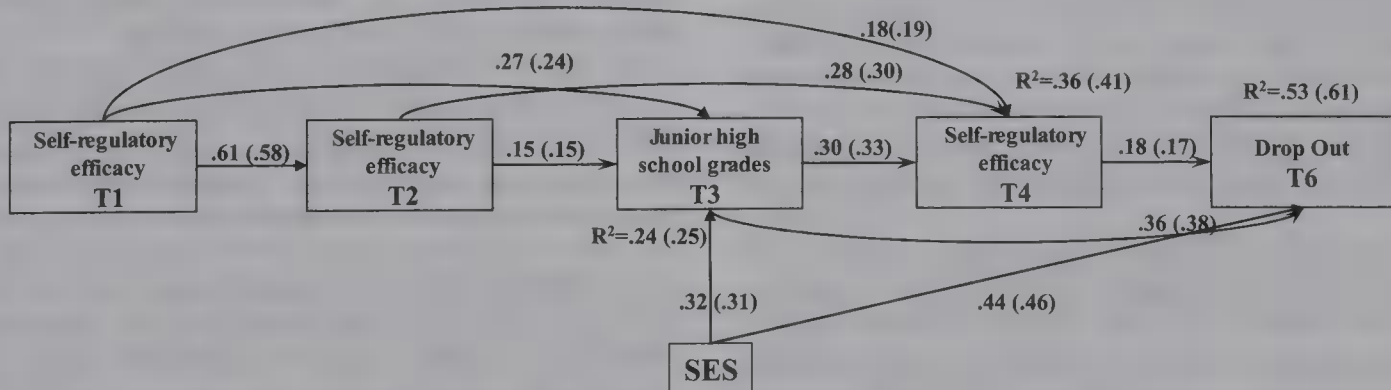


Figure 3. The role of self-regulatory efficacy in school drop-out operating in conjunction with prior academic achievement and socioeconomic status (SES). The figure includes all of the coefficients that are significant beyond the  $p = .05$  level. The first coefficient on each structural link is for males, and the second coefficient in parentheses is for females.



educational development. Thus, a high level of self-regulatory efficacy in junior high is accompanied by a lesser decline in perceived efficacy to manage learning activities with age, higher academic achievement in high school, and a lower likelihood of dropping out of school. These findings are in accord with those from other spheres of life showing that strong belief in one's personal efficacy enables people to weather taxing and stressful conditions and facilitates recovery from adverse experiences (Allen, Leadbeater, & Aber, 1990; Benight & Bandura, 2004; Lent, Brown, & Larkin, 1984).

In the longitudinal analysis, students' perceived efficacy to regulate their learning activities at the junior high level contributed to their academic achievement in high school and their likelihood of completing their high school education. Self-regulated efficacy retained its relation to academic achievement and continuance in school after we controlled for variations in prior academic performance and socioeconomic level. Academic competence is not a fixed property that one has in one's behavioral repertoire (Bandura, 1990; Sternberg & Kolligan, 1990). Rather, it is the product of ability factors and a host of self-regulatory and motivational influences (Bandura, 1997). Hence, control for prior performance not only controls for the effects of actual self-regulatory efforts but for other possible influences as well.

Findings from diverse lines of research on the contributions of self-efficacy beliefs to academic achievement further confirm that belief in one's capabilities contributes independently to academic achievement rather than simply being an ephiphenomenal reflection of prior performance (Bandura, 1997; Pajares & Schunk, 2001; Schunk & Zimmerman, 1994). The unique contribution applies equally for perceived efficacy to regulate one's learning activities (Zimmerman & Bandura, 1994; Zimmerman et al., 1992).

Advancing knowledge on academic self-development requires converging evidence from diverse methodologies because no one approach can do it alone. The research cited above sheds light on the influential role of perceived self-regulatory learning based on experimental and short-run prospective studies. The present analysis extends the analysis longitudinally. Mulaik (1987, 1993, 2001) advanced a probabilistic conception of causality that is applicable both to experimental and naturalistic studies. In his conception, causality is expressed as functional relation between variables with total independence providing the necessary condition for probabilistic causality (Mulaik, 1993). The posited conceptual model specifies the relations among the variables by fixing and constraining certain parameters in the model (Mulaik, 2007). Goodness-of-fit indices provide the means for evaluating the posited structural model. Although no single method can prove causation, verifying functional relations among factors in longitudinal analysis increase confidence in theoretically specified paths of influence.

Nor is perceived self-regulatory efficacy simply a reflection of SES. Although influenced by one's socioeconomic life conditions, perceived self-regulatory efficacy contributes independently to academic attainments and completion of high school education. These findings are in accord with those of other studies showing that the impact of SES on psychosocial functioning is, in large part, mediated through its effects on people's beliefs on their efficacy to manage their life conditions (Bandura et al., 1996, 2001; Elder, 1995; Elder, Conger, Foster, & Ardeit, 1992;

Fernández-Ballesteros, Díez-Nicolás, Caprara, Barbaranelli, & Bandura, 2002).

SES affected academic performance in high school only indirectly through its impact on prior academic attainment in junior high. However, it affected whether students continued their high school education both directly as well as mediationaly. The link between SES and academic outcomes is indirect because differences in capital can lead to variations in learning opportunities. For example, family SES may provide supportive relationships among parent and school collaborations (Coleman, 1988; Dika & Singh, 2002; Sirin, 2005). It may also create quality of educational facilities, instructional materials, teacher experience, and teacher-student ratio (Wenglinsky, 1998) that can affect success in school. Finally, lacking socioeconomic resources and performing marginally in school would create disincentives for remaining in the educational system when one is free to drop out and enter the workforce.

The findings of the present longitudinal study further demonstrate that self-regulatory efficacy can affect the course of life paths through choice processes. Occupationally relevant choices play a key role in setting the course of lifestyle trajectories (Bandura et al., 2001; Lent, Brown, & Hackett, 1994). Dropping out of school can have a widespread effect on one's future life. Many of the participants in this study have gone on to college, others have enrolled in various professional schools, and still others have entered the general workforce without further education. Some have begun to establish families. Additional assessments at the transitional phase into young adulthood should shed further light on how early self-regulatory efficacy sets in motion concatenating psychosocial changes that can eventuate in major long-term impact on one's life conditions.

Prevention of erosion of children's beliefs in their academic capabilities has greater societal implications in contemporary society than it did in the past. Decline in self-regulatory efficacy foreshadows low academic performance and school drop out. Such outcomes foreclose many options in life. In the past, youth with limited schooling had recourse to well-paying industrial and manufacturing jobs demanding minimal cognitive skills. The rapid pace of informational and technological change requires the development of cognitive competencies (Bandura, 2002). Moreover, with rapid change, knowledge and technical skills are quickly outmoded unless they are updated to fit the new occupational demands. Individuals now have to take charge of their self-development over the course of their work life.

The body of knowledge on self-regulatory processes provides guidelines for enhancing students' efficacy to manage their educational development. Some progress has been made in translating this knowledge into operational models that foster self-directedness in academic pursuits (Bandura, 1997; Pajares & Urdan, 2006; Schunk & Zimmerman, 1994; Zimmerman, 1990; Zimmerman & Cleary, 2006). Teachers and parents can teach students how to set goals, monitor their learning progress, and assess their self-efficacy for learning and self-regulation for guiding the level of motivation in ways that build their sense of efficacy for managing their academic activities. Pajares and Urdan (2006) underscored the importance to set short-term goals, foster mastery goal orientation, provide students with frequent and immediate feedback on their academic activities, assess their self efficacy for



adjusting instructional practices, and make self-regulatory practices habitual and automatic.

There are no adaptive benefits to being immobilized by self-doubts about one's capabilities and belief in the futility of effort. Cross-cultural tests of self-efficacy theory demonstrate that a resilient sense of efficacy has functional value regardless of whether one resides in an individualistic or collectivistic cultural system (Bandura, 2002; Bong, 2001; Joo et al., 2000; Lent, Brown, Nota, & Soresi, 2003). The findings of the present study, based on the Italian educational system, lend further support to the cultural generalizability of self-regulatory efficacy as well. It has the same functional value as in other educational systems.

## References

- Allen, J. P., Leadbeater, B. J., & Aber, J. L. (1990). The relationship of adolescents' expectations and values to delinquency, hard drug use, and unprotected sexual intercourse. *Development and Psychopathology*, 2, 85–98.
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.
- Bandura, A. (1990). *Multidimensional scales of perceived academic efficacy*. Stanford, CA: Stanford University.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.
- Bandura, A. (2002). Growing primacy of human agency in adaptation and change in the electronic era. *European Psychologist*, 7, 1–16.
- Bandura, A. (2006a). Adolescent development from an agentic perspective. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (Vol. 5, pp. 1–43). Greenwich, CT: Information Age.
- Bandura, A. (2006b). Toward a psychology of human agency. *Perspectives on Psychological Science*, 1, 164–180.
- Bandura, A. (2008). Reconstrual of "free will" from the agentic perspective of social cognitive theory. In J. Baer, J. C. Kaufman, & R. F. Baumeister (Eds.), *Are we free? Psychology and free will* (pp. 86–127). Oxford, England: Oxford University Press.
- Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Multifaceted impact of self-efficacy beliefs on academic functioning. *Child Development*, 67, 1206–1222.
- Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (2001). Self-efficacy beliefs as shapers of children's aspirations and career trajectories. *Child Development*, 72, 187–206.
- Benight, C. C., & Bandura, A. (2004). Social cognitive theory of post-traumatic recovery: The role of perceived self-efficacy. *Behaviour Research and Therapy*, 42, 1129–1148.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Bong, M. (2001). Role of self-efficacy and task value in predicting college students course performance and future enrollment intentions. *Contemporary Educational Psychology*, 26, 553–570.
- Bouffard-Bouchard, T. (1990). Influence of self-efficacy on performance in a cognitive task. *Journal of Social Psychology*, 130, 353–363.
- Britner, S. L., & Pajares, F. (2006). Sources of science self-efficacy beliefs of middle school students. *Journal of Research in Science Teaching*, 43, 485–499.
- Brown, A. L. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. E. Weinert & R. H. Kluwe (Eds.), *Metacognition, motivation, and learning* (pp. 65–116). Hillsdale, NJ: Erlbaum.
- Bussey, K., & Bandura, A. (1999). Social cognitive theory of gender development and differentiation. *Psychological Review*, 106, 676–713.
- Caprara, G. V. (Eds.). (2001). *La valutazione dell'autoefficacia. Interventi e contesti culturali* [Assessment of perceived self-efficacy: Intervention and cultural contexts]. Trento, Italy: Erikson.
- Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, 94, S95–S120.
- DeBowski, S., Wood, R., & Bandura, A. (2001). Impact of guided exploration and enactive exploration on self-regulatory mechanisms and information acquisition through electronic search. *Journal of Applied Psychology*, 86, 1129–1141.
- Dika, S., & Singh, K. (2002). Applications of social capital in educational literature: A critical synthesis. *Review of Educational Research*, 72(1), 31–60.
- Duncan, T. E., & Duncan, S. C. (1995). Modeling the processes of development via latent variable growth curve methodology. *Structural Equation Modeling*, 2, 187–213.
- Dweck, C. S., Davidson, W., Nelson, S., & Enna, B. (1978). Sex differences in learned helplessness: II. The contingencies of evaluative feedback in the classroom and III. An experimental analysis. *Developmental Psychology*, 14, 268–276.
- Eccles, J. S., & Midgley, C. (1989). Stage-environment fit: Developmentally appropriate classrooms for young adolescents. In C. Ames & R. Ames (Eds.), *Research on motivation in education: Goals and cognitions* (Vol. 3, pp. 139–186). New York: Academic Press.
- Elder, G. H. (1995). Life trajectories in changing societies. In A. Bandura (Ed.), *Self-efficacy in changing societies* (pp. 46–68). New York: Cambridge University Press.
- Elder, G. H., Conger, R. D., Foster, E. M., & Ardel, M. (1992). Families under economic pressure. *Journal of Family Issues*, 13, 5–37.
- Fernández-Ballesteros, R., Díez-Nicolás, J., Caprara, G. V., Barbaranelli, C., & Bandura, A. (2002). Determinants and structural relation of personal efficacy to collective efficacy. *Applied Psychology: An International Review Special Issue: Challenges of Applied Psychology for the Third Millennium*, 51, 107–125.
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, 9, 466–491.
- Furstenberg, F. F., Jr., Cook, T. D., Eccles, J., Elder, G. H., Jr., & Sameroff, A. (1999). *Managing to make it: Urban families and adolescent success*. Chicago: University of Chicago Press.
- Gore, P. A. (2006). Academic self-efficacy as a predictor of college outcomes: Two incremental validity studies. *Journal of Career Assessment*, 14, 92–115.
- Graber, J. A., Brooks-Gunn, J., & Petersen, A. C. (Eds.). (1996). *Transitions through adolescence: Interpersonal domains and context*. Mahwah, NJ: Erlbaum.
- Hansen, W., Tobler, N., & Graham, J. (1990). Attrition in substance abuse prevention research. *Evaluation Review*, 14, 677–685.
- Harter, S. (1996). Teacher and classmate influences on scholastic motivation, self-esteem, and level of voice in adolescents. In J. Juvonen & K. R. Wentzel (Eds.), *Social motivation: Understanding children's school adjustment* (pp. 11–42). Cambridge, England: Cambridge University Press.
- Istituto Italiano di Statistica. (2002). *Annuario statistico italiano 2002* [Italian yearbook of statistics 2002]. Rome, Italy: ISTAT.
- Jessor, R., Donovan, J. E., & Costa, F. M. (1991). *Beyond adolescence: Problem behavior and young adult development*. New York: Cambridge University Press.
- Joo, Y., Bong, M., & Choi, H. (2000). Self-efficacy for self-regulated learning, academic self-efficacy and internet self-efficacy in web-based instruction. *Educational Technology Research and Development*, 48, 5–17.
- Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior*, 45(1), 79–122.
- Lent, R. W., Brown, S. D., & Larkin, K. C. (1984). Relation of self-efficacy expectations to academic achievement and persistence. *Journal of Counseling Psychology*, 31, 356–362.



- Lent, R. W., Brown, S. D., Nota, L., & Soresi, S. (2003). Testing social cognitive interest and choice hypotheses across Holland types in Italian high school students. *Journal of Vocational Behavior*, 62(1), 101–118.
- McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (second ed., pp. 561–614). New York: Plenum.
- McArdle, J. J., & Anderson, E. R. (1989). Latent growth models for research on aging. In L. E. Biren & K. W. Schaie (Eds.), *The handbook of the psychology of aging* (3rd ed., pp. 21–44). San Diego, CA: Academic Press.
- Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, 55, 107–122.
- Midgley, C., Feldlaufer, H., & Eccles, J. (1989). Change in teacher efficacy and student self- and task-related beliefs in mathematics during the transition to junior high school. *Journal of Educational Psychology*, 81, 247–258.
- Mulaik, S. A. (1987). Toward a conception of causality applicable to experimentation and causal modeling. *Child Development*, 58, 18–32.
- Mulaik, S. A. (1993). Objectivity and multivariate statistics. *Multivariate Behavioral Research*, 28, 171–203.
- Mulaik, S. A. (2001). The curve-fitting problem: An objectivist view. *Philosophy of Science*, 68, 218–241.
- Mulaik, S. A. (2007). There is a place for approximate fit in structural equation modeling. *Personality and Individual Differences*, 42, 883–891.
- Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55, 463–469.
- Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide* (4th ed.) [Computer software manual]. Los Angeles, CA: Authors.
- Ogbu, J. U. (1990). *Cultural model, identity, and literacy*. New York: Cambridge University Press.
- Pajares, F., & Schunk, D. H. (2001). Self-beliefs and school success: Self-efficacy, self-concept, and school achievement. In R. Riding & S. Rayner (Eds.), *Perception* (pp. 239–266). London: Ablex.
- Pajares, F., & Urdan, T. (Eds.). (2006). *Adolescence and education: Vol. 5. Self-efficacy beliefs of adolescents*. Greenwich, CT: Information Age.
- Paris, S. G., & Newman, R. S. (1990). Developmental aspects of self-regulated learning. *Educational Psychologist Special Issue: Self-Regulated Learning and Academic Achievement*, 25, 87–102.
- Pastorelli, C., Caprara, G. V., Barbaranelli, C., Rola, J., Rozsa, S., & Bandura, A. (2001). The structure of children's perceived self-efficacy: A cross-national study. *European Journal of Psychological Assessment*, 17, 87–97.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177.
- Schunk, D. H., & Pajares, F. (2002). The development of academic self-efficacy. In A. Wigfield & J. Eccles (Eds.), *Development of achievement motivation* (pp. 16–31). San Diego, CA: Academic Press.
- Schunk, D. H., & Zimmerman, B. (Eds.). (1994). *Self-Regulation of learning and performance: Issues and educational applications*. Hillsdale, NJ: Erlbaum.
- Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, 75, 417–453.
- Sternberg, R. J., & Kolligan, J. (Eds.). (1990). *Competence considered*. New Haven, CT: Yale University Press.
- Stoolmiller, M. (1994). Antisocial behavior, delinquent peer association, and supervised wandering for boys: Growth and change from childhood to early adolescence. *Multivariate Behavioral Research*, 29, 263–288.
- Wenglinsky, H. (1998). Finance equalization and within-school equity: The relationship between education spending and the social distribution of achievement. *Educational Evaluation and Policy Analysis*, 20, 269–283.
- Wigfield, A., Eccles, J. S., Schiefele, U., Roeser, R., & Davis-Kean, P. (2006). Development of achievement motivation. In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (6th ed., pp. 933–1002). New York: Wiley.
- Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, 81, 329–339.
- Zimmerman, B. J. (1990). Self-regulating academic learning and achievement: The emergence of a social cognitive perspective. *Educational Psychology Review*, 2, 173–201.
- Zimmerman, B. J. (2000). *Attaining self-regulation: A social cognitive perspective*. San Diego, CA: Academic Press.
- Zimmerman, B. J., & Bandura, A. (1994). Impact of self-regulatory influences on writing course attainment. *American Educational Research Journal*, 31, 845–862.
- Zimmerman, B. J., Bandura, A., & Martinez-Pons, M. (1992). Self-motivation for academic attainment: The role of self-efficacy beliefs and personal goal setting. *American Educational Research Journal*, 29, 663–676.
- Zimmerman, B. J., & Cleary, T. J. (2006). Adolescents' development of personal agency: The role of self-efficacy beliefs and self-regulatory skill. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 45–69). Greenwich, CT: Information Age.
- Zimmerman, B. J., & Martinez-Pons, M. (1986). Development of a structured interview for assessing students use of self-regulated learning strategies. *American Educational Research Journal*, 23, 614–628.
- Zimmerman, B. J., & Schunk, D. H. (1989). *Self-regulated learning and academic achievement: Theory, research, and practice*. New York: Springer-Verlag.

Received January 30, 2007

Revision received November 7, 2007

Accepted November 13, 2007 ■

# Does a New Learning Environment Come Up to Students' Expectations? A Longitudinal Study

Karen D. Könings, Saskia Brand-Gruwel, and  
Jeroen J. G. van Merriënboer  
Open University of the Netherlands

Nick J. Broers  
Maastricht University

School transitions and educational innovations confront students with changes in their learning environment. Though expectations are known to influence perceptions and motivation, which, in turn, influence the effectiveness of any situation, students' expectations for a new learning environment have received little attention. This longitudinal survey, conducted with 1,335 high school students (average age, 15 years), studied students' expectations and subsequent perceptions of 5 characteristics of a new environment (fascinating content, productive learning, student autonomy, interaction, and clarity of goals) and the students' (prospective) dissatisfaction. Results showed that expectations were positively related to later perceptions. Also, high prospective dissatisfaction was related to higher actual dissatisfaction with the environment later on. Investigating expectations and prospective dissatisfaction in relation to student characteristics (i.e., motivational orientations; conceptions of learning; strategies for regulation, information processing, and affective processing) show that motivational problems and fear of failure were risk factors for educational innovations. Furthermore, students' disappointment with the new environment was related to undesirable changes in student characteristics, such as increased fear of failure. The findings stress the importance of preparing students for curricular changes.

*Keywords:* student expectations, student perceptions, educational innovation, disappointment

Students' learning environments change several times during their school career: After kindergarten, they enter primary school, followed by secondary school and, possibly, higher professional education or university. Besides this school change, students are often confronted with educational innovations in school curricula, which cause changes in school practices. Before entering a learning environment, students form expectations and build ideas about how it will be, and these expectations are known to influence subsequent perceptions (e.g., Olson, Roese, & Zanna, 1996). This is highly relevant for education, because it has been shown that students' perceptions of a learning environment are of central importance for its effects on learning (Entwistle, 1991; Entwistle & Tait, 1990). However, the role of expectations in this context has received little attention, and that is a serious omission. Students' development and their pleasure in school are likely to be disturbed when their expectations of a learning environment do not match with their later perceptions.

The current study focused on students' expectations of a new learning environment and the longitudinal effects on their subse-

quent perceptions of this environment. Students' satisfaction or dissatisfaction with the expected and perceived environment was also examined. Additionally, relations between students' expectations and prospective dissatisfaction and their learning-related characteristics were explored, as well as relations between the degree of the mismatch of expectations and later perceptions, and the development of these student characteristics.

The literature on expectations in educational contexts is broad and concerns many aspects, which, however, do not specifically deal with the expectations of a learning environment. Examples include teachers' expectations of student performances (Weinstein, 1998); students' expectations of their own performances (i.e., self-efficacy, Bandura, 1977; Lopez, Lent, Brown, & Gore, 1997); students' expectations about connections between effort and performance in relation to a positive or negative mood state (Erez & Isen, 2002); students' expectations of success in relation to task-avoidance behavior, low achievement, and dissatisfaction (Nurmi, Aunola, Salmela-Aro, & Lindroos, 2003); and students' expectations of the utility of what they are learning for their future in relation to their learning motivation (future-time perspective theory; Kauffman & Hasman, 2004). In each of these studies, clear relationships have been found between expectations and the other variables being studied.

Thus, very little research has been conducted on students' expectations with regard to characteristics of a forthcoming course or learning environment. Twenty years ago, Rosinski and Hill (1986) pointed out the importance of investigating students' expectations of the content of a course and the degree to which the course met these expectations because expectations determine the way in which students enter a course or learning environment. It has also been found that students' expectations of the course objectives

---

Karen D. Könings, Department of Psychology, Open University of the Netherlands, Heerlen, The Netherlands; Saskia Brand-Gruwel, Educational Technology Expertise Centre, Open University of the Netherlands; Jeroen J. G. van Merriënboer, Netherlands Laboratory for Lifelong Learning, Open University of the Netherlands; Nick J. Broers, Department of Psychology, Maastricht University, Maastricht, The Netherlands.

This research was supported by the Netherlands Organization for Scientific Research (NWO) under Project Number 411-01-052.

Correspondence concerning this article should be addressed to Karen Könings, Department of Psychology, Open University of the Netherlands, P.O. Box 2960, Heerlen 6401 DL, The Netherlands. E-mail: karen.konings@ou.nl



influence their perceptions of the course, even independent of what they actually encounter (Kirschner, Meester, Middelbeek, & Hermans, 1993). In spite of these results, research on students' expectations of a learning environment has lain fallow. More general psychological literature about expectations, however, indicates two reasons for taking the role of expectations in education more seriously: (a) Expectations affect the subsequent perception of a learning environment and so determine its effectiveness, and (b) expectations affect students' motivation, engagement, and investment of effort in learning.

### Expectations and Perceptions

The influence of expectations on students' perceptions of a learning environment is highly relevant because perceptions determine students' study behavior and, consequently, their performance and the effectiveness of the environment (Entwistle, 1991). Expectations can bias perceptions in three different ways. First, expectations bias information-gathering processes because they direct the learner's attention to information that is either consistent or clearly inconsistent with the expectations themselves. Both consistent and inconsistent information is more likely to be noticed, which leads to selective perception (Olson et al., 1996).

Second, expectations bias the interpretation of information because information is likely to be interpreted in a way that is consistent rather than inconsistent with expectations (Olson et al., 1996). A study on expectations of students' capabilities (Murray, 1996) has shown that ambiguous, stereotyped information about race, class, and gender affects the perceiver's estimates of the student's performances. Another example of the expectancy confirmation bias is found in diagnosing learning disabilities (Gnys, Willis, & Faust, 1995). Diagnostic decisions of school psychologists were found to be partly based on irrelevant information and false beliefs. Expectations heightened attention for illusory congruent characteristics in students' test scores and guided the interpretations and diagnoses.

Third, expectations bias subsequent behavior. People are likely to behave in a manner that is consistent with their expectations (Olson et al., 1996). A well-documented example of this phenomenon is learned helplessness (Seligman & Mayer, 1967)—the relinquishing of proactive behaviors when experiencing lack of control over the environment. Symptoms of learned helplessness have also been shown in educational contexts in which students gave up trying to perform when they did not see themselves as capable of reaching success (Craske, 1988). In addition to this direct effect on behavior, expectations may even shape the environment. People tend to behave in such a way that their behavior optimally matches their expectations, and thus, they create what they expect, a phenomenon known as a self-fulfilling prophecy (Merton, 1948). Research has shown that teachers, told that a class was highly intelligent, consequently expected higher performances, which subsequently resulted in higher student performances (the "Pygmalion in the classroom" experiment, Rosenthal & Jacobson, 1968).

Applied to education, adequate or inadequate expectations may have far-reaching effects. A student entering a learning environment with high expectations of finding certain characteristics there (e.g., student autonomy) will look for information consistent with the expectations, interpret this information in such a way that it

supports the expectations, and behave in a way that is consistent with these expectations. This student is likely to have more positive perceptions than another student entering the same environment with low expectations of autonomy because that student will mainly attend to stimuli supporting the low expectations and will interpret stimuli and behave in a way consistent with low expectations. The student with low expectations for student autonomy, consequently, will display less autonomic behavior and a more passive attitude. In contrast, the student with higher expectations is more likely to find stimuli for autonomous behavior and will tend to be more proactive. In short, students in the same learning environment are likely to perceive it differently and to behave differently, depending on their *a priori* expectations of it.

Our study focused on the effects of students' expectations of a new learning environment on their later perceptions by addressing the following research questions: (a) Do expectations of a learning environment predict how the future environment is perceived? and (b) Is students' prospective dissatisfaction associated with the extent of actual dissatisfaction they experience in a learning environment?

### Expectations, Motivation, and Learning

Investigating students' expectations is not only relevant because expectations are related to perceptions but also because research reveals that expectations affect engagement, motivation, and investment of effort. According to the expectancy-value model (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000), expectations and confidence or doubt about the attainability of the personal values influence different aspects of behavior, such as effort, persistence, and performance. These findings imply that students expecting a learning environment that corresponds with their desired environment—which means low prospective dissatisfaction—feel relatively confident and positive, which results in higher learning motivation in the future environment (see also Carver & Scheier, 2001). In contrast, students expecting a learning environment that is very different from their desired environment experience doubt and are in a negative mood, which results in low motivation and disengagement. This finding is in line with cognitive dissonance theory (Festinger, 1957), which states that inconsistencies between cognitions, feelings, and behaviors evoke a negative internal state that people try to reduce whenever possible. Cognitive dissonance is a fundamentally motivational state (Elliot & Devine, 1994), and thus, it is likely that dissonances that students experience in education will have negative motivational effects.

Consequently, a relation between expectations of a learning environment and motivation can be anticipated. More specifically, literature indicates that expectations of at least some characteristics of a learning environment can be hypothesized to be related to motivation. Contextualized and meaningful subject matter results in gains in motivation and involvement when compared with outcomes for abstract and decontextualized learning content (Cordova & Lepper, 1996). Recognizing the utility of course content leads to higher intrinsic motivation and better study habits (Simons, Dewitte, & Lens, 2004). Also, learning goals serving a directive function lead to greater investment of effort, positively affect persistence, and motivate the learner (Locke & Latham, 2002).

Besides the relation of the learning environment to motivation, students' *perceptions* of a learning environment have also been shown to be related to several other learning-related student characteristics (e.g., Luyten, Lowyck, & Tuerinckx, 2001; Wierstra & Beerends, 1996), especially to conceptions of learning (Tsai, 2000) and affective processing strategies (Könings, Brand-Gruwel, & van Merriënboer, 2005a). Motivation is only one aspect of a broader range of student characteristics that might be related to expectations. Vermunt (1996; Vermunt & Vermetten, 2004) defined five clusters of components of student learning: motivational orientations, conceptions of learning, affective processing strategies, information processing strategies, and regulation strategies. The current study explored how students' expectations relate to all components of these five clusters of learning-related characteristics. As perceptions have been shown to be related to more student characteristics than motivation alone, the same may also be true for expectations. No earlier research has focused on this aspect. Therefore, in addition to the two research questions defined earlier in *Expectations and Perceptions*, we investigated the following second set of research questions: (c) Are students' expectations of a learning environment associated with their motivational orientation and other student characteristics? and (d) Is students' prospective dissatisfaction associated with motivational orientation and other student characteristics?

### Expectations, Perceptions, and the Development of Student Characteristics

In addition to the relation between expectations and student characteristics, it is important to investigate the relation between possible differences in expectations and later perceptions (i.e., meeting expectations vs. disappointment with the environment) and the development of student characteristics. Carver and Scheier (2001) predicted strong declines in engagement in cases in which disappointment is encountered. For individuals who begin with high positivism and high engagement and who then experience situations that temper this positivism, engagement slowly decreases for a while. But at some point, a small decrease in the level of positivism produces an abrupt drop in the level of engagement.

So there are indications that the degree to which students' expectations are met influences the development of their learning-related student characteristics, like engagement or motivation. Therefore, the third set of research question is as follows: (e) Is the perceived learning environment in line with students' expectations? and (f) What is the relation between the mismatch in students' expectations and later perceptions and the developments in their learning-related characteristics?

### Powerful Learning Environments

This study investigated student expectations, perceptions, and learning-related characteristics in the context of powerful learning environments (PLEs). Such learning environments promote acquiring high-quality knowledge, problem-solving skills, self-directed learning skills, and transferability of knowledge and skills (see De Corte, Verschaffel, Entwistle, & van Merriënboer, 2003, and Könings, Brand-Gruwel, & van Merriënboer, 2005b, for an overview). Expectations with respect to five characteristics, described in the literature as fundamental to PLEs, were studied in more detail. First, PLEs should contain complex, realistic, and challenging learning tasks (van Merriënboer & Paas, 2003). Second, learning in a PLE is not directed toward reproduction of knowledge but toward an active process of making sense of the subject matter and creating mental models, which can be reused in new problem situations (Collis & Winnips, 2002; Moreno & Mayer, 1999). Third, a self-directed and independent way of learning and thinking is stimulated by gradually transferring the responsibility for the learning processes from the instructional agent to the students themselves (Vermunt, 2003). Fourth, through the inclusion of small groups, collaborative work, and ample opportunities for interaction, PLEs give students an active and constructive role in the learning process (van Merriënboer & Paas, 2003). Fifth, learning goals and task demands are clear as they direct learning strategies (Broekkamp, van Hout-Wolters, Rijlaarsdam, & van den Bergh, 2002).

The proposed concepts involved in the current study are depicted in Figure 1. At the first assessment time (T1), students' expectations of a new learning environment and their prospective

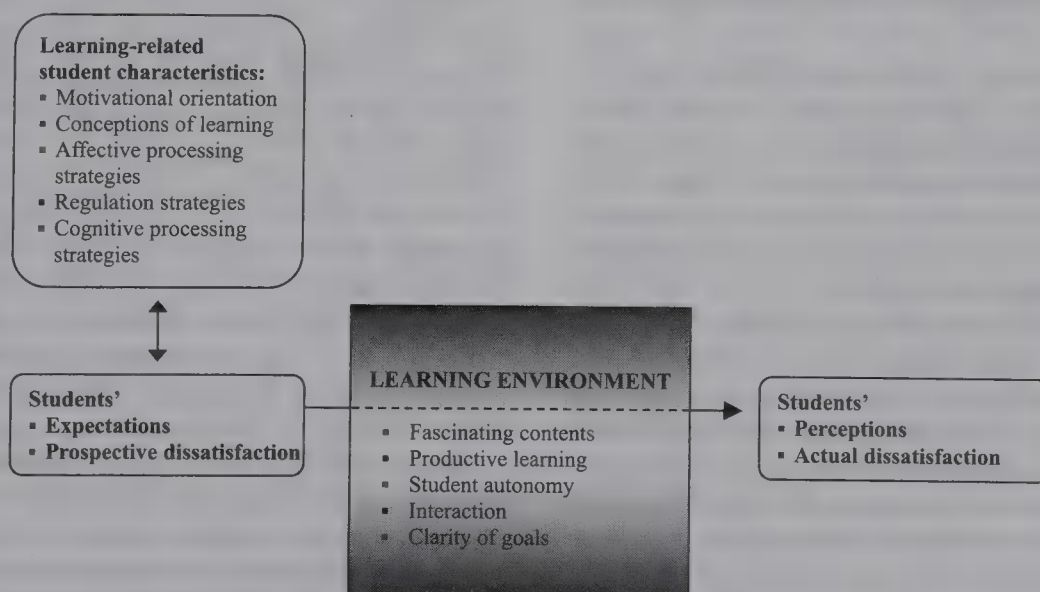


Figure 1. Visualization of the variables involved in the study.



dissatisfaction with it were measured on five characteristics of the environment. Additionally, their learning-related characteristics, like motivation, were investigated. Students reported on their perceptions and actual dissatisfaction with the learning environment after 1 year of experience (T2) and again after 2 years (T3). Relations among measures at T1, T2, and T3 were investigated as well as relations between expectations and student characteristics at T1. Finally, possible discrepancy between expectations at T1 and perceptions at T2 (i.e., disappointment) and the development of learning-related student characteristics was examined.

## Method

### Participants

At the first measurement (T1), the sample consisted of 842 students in the 9th grade (mean age = 15.27 years,  $SD = 0.52$ ) from five secondary schools in the Netherlands who were attending either senior general education (i.e., a 5-year program, preparing for higher professional education) or preuniversity education (i.e., a 6-year program, preparing for university education) classes. They were on the eve of participating in an innovative PLE in Dutch secondary education called the Second Phase. One year later (T2), the sample consisted of 1,146 student in the 10th grade, 727 of whom already had participated at T1. At T2, all students had participated in the innovative environment for 1 year. At T3, the sample consisted of 704 students in the 11th grade from four schools: 433 students participated at all three measurement moments; 181 at T2 and T3; 16 at T1 and T3, and 74 at T3 only. At T3, the 11th graders had studied in the learning environment for about 2 years. In total, 1,335 students participated in the study (50.6% girls, 49.4% boys).

The increase in the number of participants at T2 was partly due to one school's decision to allow only half of the students (i.e., classes) to participate at T1 of this study. One year later (T2), however, all 10th graders of this school participated. Furthermore, about 200 of the newly included participants at T2 were repeaters from an earlier cohort (i.e., the first year). About 20 students at T2 had been absent during data collection at T1. The attrition at T2 was likely due to incidental absence of students and nonpromotion from 9th to 10th grade.

The decrease in the number of participants at T3 was partly due to nonpromoted students who left the program and partly due to one school's decision to refrain from further data collection at T3. Because at each measurement (both at T2 and at T3), the nonpromoted students of an earlier cohort were added to the sample, a better representation of the population was created, and unwanted shifts or biases in the sample were avoided.

The achievement level of the participating schools on the general examination indicates that they are representative of schools in the Netherlands, with one school at the senior general education level scoring greatly above the national average and two schools at the preuniversity level scoring slightly above the average (Onderwijsinspectie [Dutch Inspection of Education], 2006). The percentage of students from cultural minorities at the participating schools ranged from 0.00% to 1.33% (national average is 2.55%; W. Wieldraaijer, Centraal Financiële Instellingen [Central Financial Institution], personal communication, January 8, 2007).

### Materials

*The learning environment.* The context of this study is a nationwide innovation in Dutch secondary education called the Second Phase (Ministerie van Onderwijs, Cultuur, en Wetenschap [Dutch Ministry of Education, Culture, and Science], n.d.; Stuurgroep Profiel Tweede Fase Voortgezet Onderwijs [Steering Committee for the Profile for the Second Stage of Secondary Education], 1995; Veugelers, de Jong, & Schellings, 2004). All schools in the Netherlands had to participate in this innovation. The Second Phase requires students to independently acquire skills and knowledge to better prepare them for higher professional education and university. Students learn in a self-directed way with possibilities for collaborative learning. There is more room for individual differences than in the traditional educational system, and teachers have to take these differences into account. The teacher's role is more like that of a coach and less like that of an instructor, which creates more possibilities for interaction between students and the teacher. The learning process is not only directed to knowledge acquisition but also to the selection and processing of the vast amounts of information available today. Furthermore, learning content is actualized and broadened because building a broad general knowledge base is an important goal of the Second Phase. Courses are clustered in profiles of closely interconnected topics (e.g., science and health, economics and society) that are meant to enable better integration of the subjects and lead to a better preparation for higher professional education and university. In addition, the coherence between knowledge and skills and the application of knowledge in subject-matter domains are emphasized.

The objective characteristics of the implementation of the Second Phase on the schools participating in this study are beyond the scope of this article. However, research has shown that—in general—its implementation with respect to stimulating student autonomy and differentiation are not convincingly perceived by teachers (Könings, Brand-Gruwel, & van Merriënboer, 2007). Teachers also do not perceive much freedom to deviate from lesson programs and regret this lack of freedom (Veugelers et al., 2004).

*Inventory of Expected Study Environment—Extended (IESEE) and Inventory of Perceived Study Environment—Extended (IPSEE).* The IESEE measures students' expectations of a forthcoming learning environment and their prospective dissatisfaction with regard to that environment. It is a parallel version of the IPSEE, which measures students' perceptions of a learning environment and their desires and actual dissatisfaction with regard to the design of the environment. Both IESEE and IPSEE consist of 44 items, partly based on the Inventory of Perceived Study Environment (IPSE; Wierstra, Kanselaar, van der Linden, & Lodewijks, 1999), which was translated into Dutch by the Expertise Center for Active Learning of Maastricht University (Picarelli, Slaats, Bouhuijs, & Vermunt, 2006).

The IESEE/IPSEE items are ordered in five scales (see Table 1) that can be seen as basic characteristics of PLEs. All items contain a statement about one of the characteristics of a learning environment and two statements, one related to the expectation/perception of a characteristic and one related to its desirability, as in the following example:

Table 1

*Internal Consistencies of the Scales of the Inventory of Expected Study Environment—Extended at T1 and the Inventory of Perceived Study Environment—Extended at T2 and T3*

Scale	Description of scale	No. of items	Cronbach's alpha coefficient					
			T1		T2		T3	
			Expectation	Dissatisfaction	Perception	Dissatisfaction	Perception	Dissatisfaction
Fascinating Content	Extent to which learning contents are interesting, challenging, and personally relevant for students	9	.82	.72	.85	.81	.85	.84
Productive Learning	Little emphasis on sole reproduction of learning contents but rather on an active process of making sense of the subject matter and creating mental models	5	.80	.80	.83	.84	.79	.76
Student Autonomy	Self-directedness with regard to contents and way of learning and planning	15	.81	.82	.85	.88	.84	.88
Interaction	Collaboration with peers and interaction with teacher	11	.68	.65	.73	.71	.68	.73
Clarity of Goals	Clarity of instructional goals and task demands	4	.75	.69	.81	.79	.83	.82

Note. T = time of assessment (1, 2, or 3).

Students can decide for themselves how they wish to learn during the course.

(A). I expect this to happen (in the 10th grade)/This happens.

(B). I would like this to happen.

The statements are rated on a 6-point scale, ranging from *totally disagree* (1) to *totally agree* (6). Scores on Statement A give a measure of the student's expectation/perception of the learning environment. Scores on Statement B show what the student desires from the environment. For the IESEE, the difference between the scores on Statements B and A is defined as a measure of the prospective dissatisfaction with the forthcoming environment. For the IPSEE, the difference is a measure of the actual dissatisfaction with the perceived environment. Increasing differences indicate increasing dissatisfaction. Small differences indicate low dissatisfaction. It should be noted that low dissatisfaction could be interpreted as high satisfaction, but we used only the term *dissatisfaction* to interpret and present the results in a univocal way.

**Statistics of the IESEE.** Internal consistencies of the IESEE are presented in Table 1 (T1, Columns 1 and 2). The coefficients for the expectation items ranged from .68 for the Interaction Scale to .82 for the Fascinating Content Scale. With respect to dissatisfaction scores, the alpha coefficients ranged from .65 for the Interaction Scale to .82 for the Student Autonomy Scale. In total, 3 of 10 Cronbach's alpha coefficients were above .60; 2 were above .70, and 5 were above .80. To examine whether the five scales were sufficiently independent to warrant separate consideration, we computed pairwise correlations between the scales. Of the 20 correlations (10 over expectation data and 10 over prospective dissatisfaction data), all were below .50; the implication of this finding is that

less than 25% of the variation on one scale can be explained by variation on the other scale. In addition, the tolerance was computed as a check for possible collinearity between scales. The tolerance measure, which has a range from 0 to 1, indicates serious collinearity if the values are below .10. It was computed separately for each of the five scales for perception and dissatisfaction data: The lowest value was .58, with six of the values being above .60. There was no statistical objection to consider the five IESEE-scales separately.

**Statistics of the IPSEE.** Internal consistencies were computed for all five scales for the perception items and the actual dissatisfaction scores separately at T2 and T3 (see Table 1, Columns 3–6). For the Fascinating Content Scale, the coefficients ranged between .81 and .85; for the Productive Learning Scale, between .76 and .83; for the Student Autonomy Scale, between .84 and .88; for the Interaction Scale, between .68 and .73; and for the Clarity of Goals Scale, between .79 and .83. In total, only 1 of 20 Cronbach's alpha coefficients was between .60 and .70; 6 were between .70 and .80, and another 13 were above .80. As for the IESEE-scales, correlations between the scales were computed over perception data and dissatisfaction data. At T2, 15 of 20 correlations were below .50, 4 were between .50 and .60, and 1 was above .60. The lowest tolerance value found at T2 was .45. Of the remaining values, 4 were above .50, and another 4 were above .60.

For T3, 15 correlations between the scales were below .50, and 5 were between .50 and .60. The lowest tolerance value was .49; 4 values were above .50, and another 4 were above .60. Thus, there are no statistical objections to considering the five scales separately.

**Inventory of Learning Styles for Secondary Education (ILS–SE).** The ILS questionnaire was originally developed to measure higher education students' learning styles (Vermunt, 1992) and was



adapted for students in secondary education by Vermunt, Bouhuijs, and Picarelli (2003). The questionnaire measures learning-related characteristics of students on the basis of their usual way of learning. The ILS-SE consists of 100 items. Because of the results of factor analyses, we decided to exclude a single item because of a small factor loading ( $< .40$ ). The remaining 99 items were divided in five clusters: Processing strategies (cognitive activities students use to process learning contents), regulation strategies (the way students regulate their own learning process), motivational orientations (personal goals or motives students have for learning and going to school), conceptions of learning (mental models about learning), and affective processing strategies (emotional aspects of learning). Each of the five clusters contains several scales, which are presented in Table 2.

For each item in the ILS-SE, students rate the degree to which a statement corresponds to their own learning on a 5-point scale. Information about internal consistencies of the scales at T1 and T2 is included in Table 2. At T1, Cronbach's alpha ranged from .58 for the Intrinsic Motivation and Certificate-Oriented Motivation Scales to .87 for the Motivation/Concentration Problems and Fear of Failure Scales. At T2, the coefficients ranged from .63 for the Certificate-Oriented Motivation Scale to .87 for the Fear of Failure Scale. In total, 2 of 32 Cronbach's alpha coefficients were .58, 7 were above .60, 13 were above .70, and 10 were above .80, all of which are acceptable. By computing correlations and tolerance values, we tested the independence of the 16 ILS-SE-scales. Table 3 shows the correlations between the scales. It can be seen that 116 of the correlations were below .50, 3 were between .50 and .60,

Table 2  
*Descriptions and Internal Consistencies of the Scales of the Inventory of Learning Styles for Secondary Education at T1 and T2*

Scale	Description of scale	No. of items	Cronbach's alpha coefficient	
			T1	T2
Processing strategy				
Deep processing	Relating and structuring knowledge elements and critical processing of information	12	.84	.84
Stepwise processing	Memorizing, rehearsing, studying information in detail	8	.81	.80
Regulation strategy				
Self-regulation	Regulation of own learning process through activities like planning, monitoring, reflecting, and taking initiatives with respect to learning contents	8	.71	.71
External regulation	Learning processes to be regulated by external sources (i.e., books, teacher)	6	.68	.66
Lack of regulation	Difficulties with regulating learning and processing contents effectively	4	.66	.71
Learning orientation				
Intrinsic motivation	Learning because of interest in learning content and the desire to develop oneself	4	.58	.67
Certificate oriented	Learning for passing tests, gaining high grades, and obtaining certificates	5	.58	.63
Vocation oriented	Learning for future study and professions	4	.73	.77
Ambivalent	Doubtful, uncertain attitude toward own capacities and chosen courses	5	.75	.74
Conception of learning				
Construction and use of knowledge	Learning as constructing one's own knowledge, making it concrete, and applying it	8	.82	.81
Intake of knowledge	Learning as taking in information and memorizing or reproducing it	4	.64	.64
Cooperative learning	Preferring to learn in cooperation with fellow students	3	.70	.76
Stimulating education	Learning as a process continuously driven by teachers or textbooks	5	.78	.79
Affective learning strategy				
Motivation/concentration problems	Difficulty with concentrating and staying motivated during learning, being easily distracted, and sometimes postponing assignments	8	.87	.86
Fear of failure	Experiencing stress during learning, especially in testing situations, and having a negative self-image	8	.87	.87
Keeping a good state of mind	Having a positive opinion of own capacities, being self-confident, and performing activities to stay motivated and concentrated	8	.72	.71

Note. T = time of assessment (1, 2, or 3).

Table 3

*Pearson's Correlations Between the Scales of Inventory of Learning Styles for Secondary Education at T1*

Scale	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Deep processing	—														
2. Stepwise processing	.35**	—													
3. Self-regulation	.68**	.53**	—												
4. External regulation	.28**	.54**	.39**	—											
5. Lack of regulation	.14**	.16**	.16**	.16**	—										
6. Intrinsic motivation	.37**	.27**	.39**	.18**	-.06	—									
7. Certificate oriented	.04	.35**	.14**	.33**	.07*	.05	—								
8. Vocation oriented	.17**	.21**	.22**	.22**	.03	.24**	.33**	—							
9. Ambivalent	.01	-.05	.02	-.08*	.41**	-.19**	-.11**	-.23**	—						
10. Construction and use of knowledge	.54**	.31**	.46**	.30**	-.03	.41**	.18**	.42**	-.13**	—					
11. Intake of knowledge	.01	.27**	.09**	.32**	.31**	-.01	.28**	.16**	.17**	.07	—				
12. Cooperative learning	.19**	.12**	.22**	.16**	.16**	.08*	.07	.16**	.13**	.29**	.25**	—			
13. Stimulating education	.24**	.08*	.17**	.12**	.31**	.08*	.04	.10**	.20**	.31**	.30**	.30**	—		
14. Motivation/concentration problems	-.10**	-.35**	-.29**	-.19**	.19**	-.18**	-.17**	-.15**	.23**	-.23**	.00	-.05	.12**	—	
15. Fear of failure	.22**	.19**	.31**	.09*	.48**	.09**	.03	-.01	.38**	.04	.22**	.20**	.20**	.06	—
16. Keeping a good state of mind	.37**	.32**	.37**	.33**	-.09**	.30**	.20**	.20**	-.21**	.39**	.02	.09*	.07	-.14**	-.06

\*  $p < .05$ . \*\*  $p < .01$ .

and only 1 was slightly above .60. The lowest tolerance value was .40, with 12 values above .60. Thus, there was no statistical objection to considering the 16 ILS-SE scales separately.

### Procedure

At T1, the participants filled out the IESEE and the ILS-SE. At T2, they filled out the IPSEE and the ILS-SE. At T3, they only filled out the IPSEE. Preceding the completion of a questionnaire, students received a short oral instruction about the goal and content of the questionnaire and about the way items had to be scored. This instruction was repeated on the first page of each questionnaire. The IESEE/IPSEE took between 30 and 40 min to complete; the ILS-SE took between 20 and 30 min. The participants filled out the questionnaires during regular school hours.

### Data Analysis

A maximum of 25% of missing values was accepted to compute the mean scores for each scale of the IESEE, IPSEE, and ILS-SE. If at least 75% of the items of a scale were filled out, these items were used to compute the mean score of that scale. For each scale, we could calculate a mean score using at least 95% of the participants. Dissatisfaction scores of the IESEE/IPSEE were computed as the difference between the desirability score and the expectation/perception score. To answer the research questions, we focused on students who had a positive attitude toward PLEs. This means that we analyzed only data of students who desired a particular characteristic of a learning environment to be more strongly implemented than they expected or perceived (i.e., desire - expectation/perception  $\geq 0$ ). In fact, there were also students who desired particular aspects of PLEs to be less strongly implemented, but these were rather small groups (< 10% of all students for three scales, < 20% for two scales).

Since the longitudinal design had a nested data structure, with participants nested in classes (i.e., year groups) and classes nested in schools, we expected both serial correlations due to repeated

measurements and intraclass correlations due to the multilevel structure. Data were analyzed with a longitudinal mixed model: Repeated measures were considered to be nested in participants, and participants were considered to be nested in classes. Because the number of schools was too small to permit inference to the population of schools, school was included as a fixed factor in the model to correct for correlations in the data due to nesting within schools. Thus, it is assumed that the five schools were representative of the wider population of schools. School and/or class was only included in the model if the effects was significant at a level of  $p < .10$ .

Apart from accounting for the multilevel structure of the data, the longitudinal mixed model analysis has two other advantages over traditional repeated measures analysis of variance (ANOVA). First, repeated measures ANOVA assumes that the residual variation can be described by a covariance structure known as sphericity. This is a highly restrictive assumption that is seldom realistic in the case of repeated measures. The longitudinal mixed model permits the specification of more realistic covariance structures. We opted for an unstructured covariance matrix, posing no restrictions on the values of residual variances and covariances. Second, repeated measures ANOVA discards each participant with a missing value on any of the three measurements. In contrast, the longitudinal mixed model makes use of maximum likelihood estimation. Under the assumption that cases are missing at random (MAR), participants with missing data on one or two measurements can still be used for estimation purposes. The MAR assumption is plausible in our case, so that mixed model analysis allowed for a more efficient use of the available data while still yielding unbiased estimates of effects.

A specific problem for analyzing data to answer Research Questions a, b, e, and f was the correction for class effects. Class composition changed over the time periods so that the same pupil could belong to three differently composed classes. We circumvented this problem by trying out a maximum of three different class corrections for each separate model. We first tried class as a



random factor by using the classes as composed at T1, then classes as composed at T2, and finally classes as composed at T3. In principle, this procedure could result in more significant class effects to be reported, but in practice, this did not pose a problem because the estimates of the fixed effects were unbiased, and the standard errors only marginally changed under the different class corrections. The tables that will be discussed in the Results section only report standard errors corrected for school effects (if relevant). If a significant class effect changed a parameter estimate from significant to nonsignificant or vice versa, we will explicitly discuss it in the text. In the following section, only results significant at a level of  $p < .01$  are reported.

Results

Table 4 presents the means and standard deviations of expectation scores (T1), perception scores (T2 and T3), and dissatisfaction scores (T1, T2, and T3).

Relationships Between Students' Reports on Different Measurement Moments

To investigate how expectations of a learning environment predict perceptions of the future environment and how prospective dissatisfaction predicts actual dissatisfaction (i.e., Research Questions a and b), we conducted mixed model regression analyses to examine mutual relations between expectation scores at T1, perception scores at T2, and perception scores at T3. To investigate the relation between expectation scores at T1 and perception scores at T2, we tested a model with only data from T1 and T2: the perception score at T2 as a dependent variable and the expectation score at T1 as an independent variable. We examined the relations between expectation scores (at T1) and perception scores at T2 and T3 by building a model with perception scores at T3 as dependent variable and expectation scores at T1 and perception scores at T2 as independent variables. Testing this model provided insight in the relation between scores at T1 and T3 and between scores at T2 and T3. Because perception scores at both T2 and T3 were included in the model, the regression coefficient of the score at T2 was corrected for the score at T3 and vice versa. The regression coefficients represent the size of the unique part of the relation between the dependent and independent variable.

*Expectation and perception scores.* The left side of Table 5 presents the results of analyzing the mutual relations between expectation scores at T1 and perception scores at T2 and T3. The expectation scores at T1 had a significant positive effect on perception scores at T2 for all scales. Thus, the higher the expectation scores, the higher the perception scores at T2. Perception scores at T2 also had a significant positive effect on perception scores at T3. But as can be seen from Table 5, the direct effect of expectation scores (T1) on perception scores at T3 was nonsignificant for two scales and relatively small for the other scales. This result is likely due to the mediating role of the perception scores at T2. By including perception scores at T2 in the analyses, we corrected the results for this potential mediator and showed the size of the unique relation between scores at T1 and T3.

*Dissatisfaction scores.* Results for relations between dissatisfaction scores at T1, T2, and T3 (see the right side of Table 5) show that, for all scales, prospective dissatisfaction at T1 had a positive effect on actual dissatisfaction at T2, and dissatisfaction at T2 had a positive effect on dissatisfaction at T3. Prospective dissatisfaction at T1 had a direct positive effect on actual dissatisfaction scores at T3 on two of the scales, indicating a unique relation between the dissatisfaction scores at T1 and T3 for Fascinating Content and Clarity of Goals Scales.

In summary, the results for Research Questions a and b show robust relations between expectations and later perceptions. The higher students' expectations before entering the new learning environment, the higher their subsequent perceptions later on. Prospective dissatisfaction scores were positively related to actual dissatisfaction scores with the perceived learning environment.

Relationships Between Students' Reports at T1 and Learning-Related Student Characteristics

To investigate how expectations of the future learning environment are related to motivation and other learning-related student characteristics and how prospective dissatisfaction is related to these student characteristics (i.e., Research Questions c and d), we conducted mixed model regression analyses to analyze the relations between IESEE scores at T1 and learning-related student characteristics at T1. The learning-related student characteristics were included as independent variables in mixed model regression analyses. A backward procedure was used, in which the less

Table 4  
Means and Standard Deviations of Expectation and Perception Scores and Dissatisfaction Scores

Scale	Expectation (T1)/perception (T2 & T3) score						Dissatisfaction score					
	T1		T2		T3		T1		T2		T3	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Fascinating Content	3.46	0.82	3.10	0.85	3.12	0.86	1.35	0.80	1.77	0.96	1.82	0.96
Productive Learning	2.76	0.97	2.89	1.01	3.32	0.93	1.53	1.02	1.29	1.00	0.93	0.81
Student Autonomy	3.88	0.70	3.29	0.71	3.38	0.69	1.11	0.74	1.37	0.88	1.32	0.86
Interaction	3.98	0.64	3.70	0.65	3.74	0.62	0.77	0.57	0.94	0.64	0.92	0.64
Clarity of Goals	4.12	0.94	3.82	0.96	3.88	1.00	1.32	0.95	1.56	1.04	1.47	1.10

Note. Scales were from the Inventory of Expected Study Environment—Extended and the Inventory of Perceived Study Environment—Extended. T = time of assessment (1, 2, or 3).

Table 5  
*Mixed Regression Coefficients of Scores for Expectation/Perception and for Prospective Dissatisfaction and Actual Dissatisfaction with the Perceived Learning Environment*

Scale	Expectation/perception score			Dissatisfaction score		
	<i>B</i>	<i>SE B</i>	$\beta$	<i>B</i>	<i>SE B</i>	$\beta$
T1-T2						
Fascinating Content	.46	.03	.44	.57	.06	.40
Productive Learning	.36 <sup>a</sup>	.04	.34	.47	.06	.40
Student Autonomy	.36 <sup>a</sup>	.04	.36	.40 <sup>a</sup>	.06	.38
Interaction	.39 <sup>a</sup>	.04	.38	.34 <sup>a</sup>	.06	.36
Clarity of Goals	.43 <sup>a</sup>	.03	.42	.48 <sup>a</sup>	.05	.41
T2-T3						
Fascinating Content	.48	.05	.48	.54 <sup>a</sup>	.05	.54
Productive Learning	.34 <sup>a</sup>	.04	.37	.31	.05	.38
Student Autonomy	.45 <sup>a</sup>	.05	.45	.56 <sup>a</sup>	.05	.59
Interaction	.52	.04	.55	.52 <sup>a</sup>	.05	.53
Clarity of Goals	.56	.05	.54	.55	.05	.54
T1-T3						
Fascinating Content	.21	.05	.20	.18 <sup>a</sup>	.06	.15
Productive Learning	.23 <sup>a</sup>	.04	.24	<i>ns</i>		
Student Autonomy	<i>ns</i>			<i>ns</i>		
Interaction	<i>ns</i>			<i>ns</i>		
Clarity of Goals	.13	.05	.12	.21	.05	.19

Note. Coefficients:  $p < .01$ . Scales were from the Inventory of Expected Study Environment—Extended and the Inventory of Perceived Study Environment—Extended. Standard errors were based on estimation of fixed effects without correction for class effects but with correction for school effects (if  $p_{\text{school}} < .10$ ). Additional correction for class effects did not change the significance of the result, unless stated in text. T = time of assessment (1, 2, or 3).

<sup>a</sup>Corrected for school effects.

significant variables were removed one by one, until all included variables were significant ( $p < .01$ ).

**Expectation Scores and Learning-Related Student Characteristics.** Table 6 presents the results on relations between expectation scores and learning-related student characteristics. As can be seen from this table, some student characteristics were related to expectation scores on several scales. These will be described in more detail. Intrinsic motivation was positively related to students' expectation scores for the Fascinating Content, Student Autonomy, Interaction, and Clarity of Goals Scales. The conception of learning as the construction and use of knowledge was also frequently related to high expectation scores: The stronger this conception, the higher students' expectations were with respect to Fascinating Content and Student Autonomy Scales. Furthermore, the use of external regulation strategies was positively related to expectation scores for Interaction and Clarity of Goals Scales. Finally, the report of fear of failure was negatively related to expectation scores for Student Autonomy and Clarity of Goals Scales. Thus, the higher the reported fear of failure was, the lower the expectations on these scales.

With respect to the formulated hypotheses about the relation between motivation and expectations for the Fascinating Content and Clarity of Goals Scales, the results show that motivational orientations were, for those scales, indeed related to expectations. For both scales, intrinsic motivation was related to holding higher expectations on how the future learning environment would look.

Expectations for Fascinating Content were negatively related to problems with motivation and concentration, while the ambivalent motivational orientation especially was negatively related to expectations about Clarity of Goals. As can be seen from Table 6, expectation scores for both scales were also related to some other learning-related student characteristics.

In summary, the results show that learning-related student characteristics were related to expectations of the future environment and, mostly, in a consistent fashion related to either higher or lower expectation scores. Especially, students reporting fear of failure tended to expect a less powerful learning environment, while students with an intrinsic motivational orientation and students with a constructivist conception of learning tended to expect the future environment to be a more powerful one.

**Prospective Dissatisfaction Scores and Learning-Related Student Characteristics.** Table 7 presents results of mixed model regression analyses on the prospective dissatisfaction data. Some of the student characteristics were often related to dissatisfaction scores, either positively or negatively. An intrinsic motivational orientation was negatively related to prospective dissatisfaction scores for four scales. The stronger students' intrinsic motivation for learning was, the lower their prospective dissatisfaction scores on the Fascinating Content, Productive Learning, Student Autonomy, and Clarity of Goals Scales. Also, the conception of learning as construction and use of knowledge was related to low prospec-



Table 6  
Significant Results of Mixed Model Analyses, Showing Relations Between Expectations and Learning-Related Student Characteristics

Dependent variable/scale	Independent variable	B	SE B	$\beta$
Fascinating Content	Intrinsic motivation	.39	.04	.31
	Construction and use of knowledge	.31	.04	.27
	Motivation/concentration problems	-.12	.03	-.14
Productive Learning <sup>a</sup>	Intake of knowledge	-.22	.04	-.16
	Motivation/concentration problems	-.12	.04	-.12
	Stepwise processing	-.16	.05	-.13
	Deep processing	.15	.06	.10
Student Autonomy <sup>a</sup>	Construction and use of knowledge	.18	.04	.20
	Fear of failure	-.11	.03	-.11
	Intrinsic motivation	.11	.04	.10
Interaction <sup>a</sup>	External regulation	.16	.03	.18
	Intrinsic motivation	.14	.03	.15
Clarity of Goals <sup>a</sup>	Ambivalent	-.19	.05	-.17
	External regulation	.18	.05	.14
	Intrinsic motivation	.18	.05	.12
	Fear of failure	-.13	.05	-.09

Note. Significant results:  $p < .01$ . Scales were from the Inventory of Expected Study Environment—Extended and the Inventory of Perceived Study Environment—Extended. Standard errors were based on estimation of fixed effects without correction for class effects, but with correction for school effects (if  $p_{\text{school}} < .10$ ). Additional correction for class effects did not change the significance of the result, unless stated in text.

<sup>a</sup>Corrected for school effects.

tive dissatisfaction, especially for Productive Learning and Interaction Scales.

Positively related to prospective dissatisfaction was an ambivalent motivational orientation: The stronger students' ambivalent motivation, the higher their prospective dissatisfaction scores for the Fascinating Content, Interaction, and Clarity of Goals Scales. Problems with motivation and concentration were related to higher prospective dissatisfaction scores for Fascinating Content, Productive Learning, and Student Autonomy Scales. Another three student characteristics were also related to higher prospective dissatisfaction scores: the conception of learning as intake of knowledge, the certificate-learning orientation, and the affective strategy keeping good state of mind (each for two IESEE scales).

The analyses of prospective dissatisfaction scores and learning-related student characteristics reveal that some student characteristics, and particularly intrinsic motivation, are related to low prospective dissatisfaction scores. Other student characteristics, such as an ambivalent motivational orientation and problems with motivation and concentration, are related to high prospective dissatisfaction scores.

Summarizing, results for Research Questions c and d show that learning-related student characteristics are related to expectations. Fear of failure was frequently related to lower expectations, whereas intrinsic motivation and a conception of learning as the construction and use of knowledge were related to higher expectations. Prospective dissatisfaction was often related to an ambivalent motivational orientation and problems with motivation and concentration. Intrinsic motivation was frequently related to low prospective dissatisfaction.

### Mismatch Between Expectations and Perceptions and Development of Student Characteristics

To examine whether students' perceptions of an environment are in line with their expectations of it (i.e., Research Question e), we used a longitudinal mixed model analysis. For testing longitudinal effects over time,  $F$  values were computed, and for identifying the exact differences among the three times of measurement, pairwise comparisons with Bonferroni correction were conducted, and Cohen's  $d$  effect size was computed. Only differences with  $d > .20$  are described in the text.

For investigating the relation between the size of the mismatch between expectations and later perceptions and the development of learning-related student characteristics in the same period (i.e., Research Question f), we conducted mixed model regression analyses in the same way as the analyses for Research Questions c and d. The mismatch between expectation scores at T1 and perception scores at T2 ( $T2 - T1$ ) ranged from  $-5$ , indicating a large decrease in scores from T1 to T2 and strong disappointment, to  $+5$ , indicating a large increase in scores from T1 to T2 and thus much higher perceptions than previously expected. Changes in each learning-related student characteristic ( $T2 - T1$ ) were included as independent variables in the model. By using a backward procedure, we built a model that only contained variables that were significant at  $p < .01$  ( $p_{\text{school/class}} < .10$ ).

**Expectation and Perception Scores.** Table 8 shows the results of the mixed model longitudinal analyses on expectation and perception scores. The results of the  $F$  tests show that significant longitudinal effects existed on all scales of the IPSEE ( $p < .01$ ). The differences between expectation scores (T1) and perception scores (T2) show that the scores decreased on four of the five

Table 7  
Significant Results of Mixed Model Analyses, Showing Relations Between Prospective Dissatisfaction and Learning-Related Student Characteristics

Dependent variable/scale	Independent variable	B	SE B	$\beta$
Fascinating Content <sup>a</sup>	Intrinsic motivation	-.28	.04	-.23
	Certificate oriented	.23	.05	.16
	Motivation/concentration problems	.15	.03	.16
	Ambivalent	.14	.04	.13
Productive Learning	Motivation/concentration problems	.21	.04	.15
	Intrinsic motivation	-.21	.06	-.12
	Construction and use of knowledge	-.22	.06	-.12
	Intake of knowledge	.13	.05	.08
	Keeping good state of mind	.15	.06	.09
Student Autonomy	Motivation/concentration problems	.12	.03	.13
	Certificate oriented	.15	.05	.10
	Intrinsic motivation	-.11	.04	-.09
Interaction <sup>a</sup>	Cooperative learning	.13	.03	.16
	Construction and use of knowledge	-.19	.04	-.19
	Deep processing	.14	.04	.13
	Ambivalent	.09	.03	.10
	Keeping good state of mind	.10	.03	.10
Clarity of Goals	Intrinsic motivation	-.20	.05	-.14
	Fear of failure	.15	.05	.13
	Ambivalent	.14	.05	.12
	Intake of knowledge	.12	.04	.09

Note. Significant results:  $p < .01$ . Scales were from the Inventory of Expected Study Environment—Extended and the Inventory of Perceived Study Environment—Extended. Standard errors were based on estimation of fixed effects without correction for class effects, but with correction for school effects (if  $p_{\text{school}} < .10$ ). Additional correction for class effects did not change the significance of the result, unless stated in text.

<sup>a</sup>Corrected for school effects.

scales, indicating disappointing perceptions compared with the expectations. The effect size was large for the Student Autonomy Scale (.84). For most scales, the differences between perception scores at T2 and T3 showed no significance. An increase of perception scores from T2 to T3 was found only for the Productive Learning Scale. Apparently, students perceived this element of the environment as being present more strongly at T3 than at T2. Scores on the Productive Learning Scale, notably, increased year after year (from T1 to T2, and from T2 to T3). However, the most striking result is the large decline of expectation scores at T1 and perception scores at T2 on the majority of the scales. Apparently,

the perceived learning environment did not meet students' expectations.

*Relations Between Disappointment and Development of Student Characteristics.* Results revealed that a decrease from expectation scores (T1) to perception scores (T2) on the Fascinating Content Scale was related to a decrease in intrinsic motivation ( $B = .24$ ;  $SE B = .05$ ;  $\beta = .20$ ), a decrease in reported use of deep processing strategies ( $B = .20$ ;  $SE B = .05$ ;  $\beta = .13$ ), and an increase in fear of failure ( $B = -.21$ ;  $SE B = .04$ ;  $\beta = -.18$ ) from T1 to T2. Thus, the larger the disappointment, the more intrinsic motivation and use of deep processing strategies decreased, and

Table 8  
Results of Mixed Model Analyses on Longitudinal Data of Expectations (T1) and Perceptions (T2 and T3)

Scale	F	df	T2-T1			T3-T2		
			$\Delta$	SE	d	$\Delta$	SE	d
Fascinating Content	76.38*	2, 368.59	-.36*	.03	.43	-.02	.03	.02
Productive Learning	98.72*	2, 618.65	.13*	.04	.13	.41*	.04	.42
Student Autonomy	243.21*	2, 582.42	-.59*	.03	.84	.07	.03	.10
Interaction	65.90*	2, 501.33	-.28*	.03	.44	.05	.02	.08
Clarity of Goals	45.80*	2, 428.35	-.33*	.04	.35	.02	.04	.02

Note. Scales were from the Inventory of Expected Study Environment—Extended and the Inventory of Perceived Study Environment—Extended. Standard errors were based on estimation of fixed effects without correction for class effects, but with correction for school effects (if  $p_{\text{school}} < .10$ ). Additional correction for class effects did not change the significance of the result, unless stated in text. T = time of assessment (1, 2, or 3).

\*  $p < .01$ .



the more fear of failure increased. For the Student Autonomy Scale, disappointment was related to an increase in fear of failure ( $B = -.15$ ;  $SE B = .04$ ;  $\beta = -.14$ ) and a decrease in deep processing ( $B = .16$ ;  $SE B = .05$ ;  $\beta = .11$ ). For the Interaction Scale, disappointment was related to a decrease in deep processing strategies ( $B = .16$ ;  $SE B = .05$ ;  $\beta = .13$ ), a decrease in considering learning as a cooperative activity ( $B = .08$ ;  $SE B = .03$ ;  $\beta = -.11$ ), and an increase in the ambivalent motivational orientation ( $B = -.10$ ;  $SE B = .04$ ;  $\beta = .10$ ). For the Clarity of Goals Scale, disappointment was related to a decrease in intrinsic motivation ( $B = .18$ ;  $SE B = .05$ ;  $\beta = .13$ ) and in keeping a good state of mind ( $B = .20$ ;  $SE B = .06$ ;  $\beta = .13$ ). Productive Learning was the only scale showing an increase in scores from T1 to T2. Exceeding the expectations for the Productive Learning Scale was related to a decrease in the conception of learning as intake of knowledge ( $B = -.15$ ;  $SE B = .05$ ;  $\beta = -.10$ ) and a decrease in fear of failure ( $B = -.15$ ;  $SE B = .06$ ;  $\beta = -.10$ ).

Summarizing, results for Research Questions e and f showed that students' perceptions of the new learning environment on most scales did not meet their expectations and that disappointment was related to undesirable changes in learning-related student characteristics, especially to an increase in fear of failure, a decrease in deep processing, and a decrease in intrinsic motivation.

### Discussion

The current study aimed to shed light on the role of expectations in education, especially expectations students have of a future learning environment. Because students move to new learning environments several times during their school career, it is relevant to gain insight in their expectations and subsequent perceptions of these new environments and to investigate how learning-related student characteristics relate to those expectations.

Research Questions a and b focused on the relations between expectations of how the future learning environment would look and later perceptions and relations between prospective dissatisfaction and actual dissatisfaction while perceiving the environment. Expectations were positively related to subsequent perceptions of the environment for all measured aspects of the environment. Thus, the higher the expectations beforehand, the higher the perceptions later on, and the lower the expectations beforehand, the lower the perceptions later on. Perceptions measured at the second and third time were also clearly positively related. Additionally, prospective dissatisfaction with the new environment was related to actual dissatisfaction after 1 year, and dissatisfaction after 1 year of participating in the environment was related to dissatisfaction after 2 years.

The relations between expectations/perceptions and dissatisfaction over time may well be explained by the cognitive biases described in the introduction of this article (see, e.g., Olson et al., 1996). People selectively pay attention to information consistent with their expectations and also interpret this information in such a way that their expectations are confirmed. Additionally, people create self-fulfilling prophecies because they behave in agreement with their expectations.

For Research Questions c and d, relations between expectations and learning-related student characteristics were investigated, as well as the way prospective dissatisfaction was related to those characteristics. Having an intrinsic motivational orientation and

conceiving of learning as the construction and use of knowledge are both related to higher expectations of the new environment. Reporting fear of failure frequently relates to low expectations; thus, students who report a strong fear of failure are more reserved in their expectations of the future environment. These findings confirm the assumed relation between motivation and expectations but also show that expectations are influenced by conceptions of learning and affective processing strategies.

Prospective dissatisfaction with the new learning environment is negatively related to intrinsic motivation for learning; that is, intrinsically motivated students think they will be satisfied with the new environment. On the contrary, motivation/concentration problems and an ambivalent motivation are strongly related with high prospective dissatisfaction; students with these characteristics think they will be unhappy with the new environment.

The finding that high expectations are related to high intrinsic motivation and an active view on learning is in agreement with the relation between positivism and "engagement," proposed by Carver and Scheier (2001). Low expectations are especially related to fear of failure. It would be an oversimplification to consider this as a form of "low engagement." Students with fear of failure prefer a high degree of structure, clearness, stability, and continuity in their learning environment (Hermans, 1975). They are averse to unexpected and unfamiliar situations. Low expectations of the new and thus unfamiliar learning environment are better understandable in this context.

The finding that high prospective dissatisfaction with the new environment is, among other things, related to problems with motivation and concentration and to an ambivalent motivational orientation fits in the expectancy-value model (Carver & Scheier, 2001; Eccles & Wigfield, 2002). That model holds that discrepancies between expectations and personal values influence persistence and the amount of effort invested in learning and may induce a sense of doubt or negative thinking. Ambivalent motivation and problems with motivation and concentration are clear signals of doubt and negativism, and they thus may indicate a lack of persistence and unwillingness to invest effort in learning. The finding that intrinsic motivation for learning relates to low prospective dissatisfaction is in agreement with results on positive thinking and feeling confident, as described in the literature. Literature about expectations and motivation proposes causality between expectations and motivation, but it is well imaginable that the effect is bidirectional. A well-motivated student may recognize the utility of course contents more than a less-motivated student. Since the main focus of our study was trying to understand student expectations, we have focused on expectations as dependent variables and investigated whether there are more (or other) student characteristics than motivation that relate to students' expectations and prospective dissatisfaction. We could only test this by including student characteristics as independent variables, to be related to expectations/prospective dissatisfaction on a particular characteristic of the learning environment. In this way, it became clear that, besides motivation, fear of failure and certain conceptions of learning also relate to expectations.

Research Questions e and f concerned possible discrepancies between expectations of the new environment and later perceptions of it. We sought to determine whether students' expectations of an innovative learning environment in Dutch secondary education were met (i.e., Research Question e). Results clearly



show that students' perceptions of the new environment fell short of their expectations. Expectations were higher than the perceptions after 1 year with respect to Fascinating Content, Student Autonomy, Interaction, and Clarity of Goals Scales. The Productive Learning Scale was the only aspect for which perceptions exceeded the expectations after 1 year and for which perceptions increased even further in the second year. This is a positive finding, because the innovative environment indeed aimed to stimulate active processing and application of knowledge, rather than reproductive learning. However, the disappointing perceptions of the other aspects of the environment are worrying. Apparently, the Second Phase is implemented in such a way that students do not perceive its valuable aspects as much as they had expected beforehand. This is problematic because perceptions direct students' learning behavior (Entwistle, 1991), which eventually determine whether educational goals of the environment will be reached.

Results on relations concerning the mismatch between expectations and perceptions and the developments in student characteristics (i.e., Research Question f) indicate that a mismatch is related to negative changes in student characteristics. Disappointment is related to increasing fear of failure, lower intrinsic motivation, and less use of deep processing strategies. Negative effects of disappointment were proposed by Carver and Scheier (2001) and by the expectancy-value model (Eccles & Wigfield, 2002), but our results further refine the nature of these effects.

A first theoretical implication of our findings is that principles from general psychological research on expectations are also applicable to an educational setting, in particular, a setting in which students are confronted with the implementation of a new learning environment. A second implication is that the concept of expectations deserves a much more prominent place in educational research than it has today. As shown in our study, students' expectations of a new learning environment are not automatically in line with later perceptions, and, even more important, expectations strongly influence the way they perceive the environment after it has been implemented. Perceptions are likely to determine students' learning behaviors and, consequently, the effectiveness of the learning environment. Gaining more insight into the role of student expectations is thus of utmost importance in developing guidelines for the design of PLEs, preferably in such a way that students' expectations are taken into account.

A practical implication of our findings is that schools and teachers should carefully prepare their students for curricular changes or innovations. The quality and quantity of information that students receive on the characteristics of a new learning environment before they start to work in it should be carefully determined to help them build proper expectations. If at all possible, disappointing perceptions should be prevented. Students with a high fear of failure are particularly vulnerable in situations of change: They should be given extra support and structure before and during the implementation of a new environment.

A limitation of the current study is that students were always forced to report their expectations of the environment in the questionnaire, regardless of the clearness of their expectations, and it is unknown how students formed their expectations and which sources of information they used to form them. Future research should focus on the origin of the expectations, including the sources that students use to form them (e.g., press, siblings, peers,

parents), in order to gain more insight into the processes yielding the expectations that students reported in this study. This can provide more insight into the nature of expectations in an educational setting and might help researchers to develop a theory of how expectations can best be dealt with in educational design. Furthermore, knowledge about the origin of expectations would provide schools with valuable information that they could use to optimize their preparation of students for curricular changes and innovations. In line with this, future research should also address the question of how the process of forming expectations could be influenced to result in more accurate expectations, that is, expectations that match later perceptions.

To conclude, this study showed that expectations of a learning environment deserve a prominent role in educational research and praxis. Students do not automatically form proper expectations of a new environment such as the Second Phase in Dutch secondary education. Nevertheless, these expectations influence their perceptions of the new environment. Disappointing perceptions are likely to decrease the effectiveness of the environment and are also related to undesirable changes in learning-related student characteristics. More effective approaches are needed to prepare students for large educational changes; such approaches should also take into account differences in individual learning characteristics and related prospective ideas. It would be highly beneficial for educational design if guidelines were developed that took into account students' expectations of curricular changes or innovations.

## References

- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84, 195–215.
- Broekkamp, H., van Hout-Wolters, B. H. A. M., Rijlaarsdam, G., & van den Bergh, H. (2002). Importance in instructional text: Teachers' and students' perceptions of task demands. *Journal of Educational Psychology*, 94, 260–271.
- Carver, C. S., & Scheier, M. F. (2001). Optimism, pessimism, and self-regulation. In E. C. Chang (Ed.), *Optimism and pessimism: Implications for theory, research, and practice* (pp. 31–51). Washington, DC: American Psychological Association.
- Collis, B., & Winnips, K. (2002). Two scenarios for productive learning environments in the workplace. *British Journal of Educational Technology*, 33, 133–148.
- Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, 88, 715–730.
- Craske, M. (1988). Learned helplessness, self-worth motivation, and attribution retraining for primary school children. *British Journal of Educational Psychology*, 58, 152–164.
- De Corte, E., Verschaffel, L., Entwistle, N., & van Merriënboer, J. J. G. (Eds.) (2003). *Powerful learning environments: Unravelling basic components and dimensions*. Oxford, UK: Elsevier Science.
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53, 103–132.
- Elliot, A. J., & Devine, P. G. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology*, 67, 382–394.
- Entwistle, N. J. (1991). Approaches to learning and perceptions of the learning environment. *Higher Education*, 22, 201–204.
- Entwistle, N. J., & Tait, H. (1990). Approaches to learning, evaluations of teaching, and preferences for contrasting academic environments. *Higher Education*, 19, 169–194.
- Erez, A., & Isen, A. M. (2002). The influence of positive affect on the



- components of expectancy motivation. *Journal of Applied Psychology*, 87, 1055–1067.
- Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.
- Gnys, J. A., Willis, W. G., & Faust, D. (1995). School psychologists' diagnoses of learning disabilities: A study of illusory correlation. *Journal of School Psychology*, 33, 59–73.
- Hermans, H. J. M. (1975). *Prestatiemotief en faalangst in gezin en onderwijs* [Performance motive and fear of failure in family and education]. Amsterdam, The Netherlands: Swets & Zeitlinger.
- Könings, K. D., Brand-Gruwel, S., & van Merriënboer, J. J. G. (2005a, August). *Students' perspective on an innovative learning environment and the relation with learning-related student characteristics*. Paper presented at the European Association for Research on Learning and Instruction conference, Cyprus, Greece.
- Könings, K. D., Brand-Gruwel, S., & van Merriënboer, J. J. G. (2005b). Towards more powerful learning environments through combining the perspectives of designers, teachers and students. *British Journal of Educational Psychology*, 75, 645–660.
- Könings, K. D., Brand-Gruwel, S., & van Merriënboer, J. J. G. (2007). Teachers' perspective on innovations: Implications for educational design. *Teaching and Teacher Education*, 23, 985–997.
- Kauffman, D. F., & Husman, J. (2004). Effects of time perspective on student motivation: Introduction to a special issue. *Educational Psychology Review*, 16, 1–7.
- Kirschner, P. A., Meester, M., Middelbeek, E., & Hermans, H. (1993). Agreement between student expectations, experiences and actual objectives of practicals in the natural sciences at the Open University of The Netherlands. *International Journal of Science Education*, 15, 175–197.
- Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year Odyssey. *American Psychologist*, 57, 705–717.
- Lopez, F. G., Lent, R. W., Brown, S. D., & Gore, P. A. (1997). Role of social-cognitive expectations in high school students' mathematics-related interest and performance. *Journal of Counseling Psychology*, 44, 44–52.
- Luyten, L., Lowyck, J., & Tuerlinckx, F. (2001). Task perception as a mediating variable: A contribution to the validation of instructional knowledge. *British Journal of Educational Psychology*, 71, 203–223.
- Merton, R. K. (1948). The self-fulfilling prophecy. *Antioch Review*, 8, 193–210.
- Ministerie van Onderwijs, Cultuur, en Wetenschap [Dutch Ministry of Education, Culture, and Science]. (n.d.) *Dossier Tweede Fase* [Dossier Second Phase]. Retrieved August 1, 2005, from <http://www.minocw.nl/tweedefase/factsheet.html>
- Moreno, R., & Mayer, R. E. (1999). Multimedia-supported metaphors for meaning making in mathematics. *Cognition and Instruction*, 17, 215–248.
- Murray, C. B. (1996). Estimating achievement performance: A confirmation bias. *Journal of Black Psychology*, 22, 67–85.
- Nurmi, J. E., Aunola, K., Salmela-Aro, K., & Lindroos, M. (2003). The role of success expectation and task-avoidance in academic performance and satisfaction: Three studies on antecedents, consequences and correlates. *Contemporary Educational Psychology*, 28, 59–90.
- Olson, J. M., Roese, N. J., & Zanna, M. P. (1996). Expectancies. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 211–238). New York: Guilford Press.
- Onderwijsinspectie [Dutch Inspection of Education]. (2006). *Opbrengstenkaarten* [Achievement cards]. Retrieved December 15, 2006, from <http://www.onderwijsinspectie.nl/>
- Picarelli, A., Slaats, M., Bouhuijs, P. A. J., & Vermunt, J. D. (2006). *Leerstijl en leeromgeving in het Voortgezet Onderwijs: Nederland en Vlaanderen vergeleken* [Learning style and learning environment in secondary education: The Netherlands and Flanders compared]. *Pedagogische Studien*, 83, 139–155.
- Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom*. New York: Holt, Rinehart, and Winston.
- Rosinski, E. F., & Hill, P. M. (1986). Student expectations and personal perceptions as an approach to course evaluation. *Medical Education*, 20, 228–233.
- Seligman, M. E. P., & Mayer, S. F. (1967). Failure to escape traumatic shock. *Journal of Experimental Psychology*, 74, 1–9.
- Simons, J., Dewitte, S., & Lens, W. (2004). The role of different types of instrumentality in motivation, study strategies, and performance: Know why you learn, so you'll know what you learn! *British Journal of Educational Psychology*, 74, 343–360.
- Stuurgroep Profiel Tweede Fase Voortgezet Onderwijs. [Steering Committee for the Profile for the Second Stage of Secondary Education]. (1995). *Verschil moet er zijn* [There should be a difference]. The Hague, The Netherlands: Author.
- Tsai, C. C. (2000). Relationships between student scientific epistemological beliefs and perceptions of constructivist learning environments. *Educational Research*, 42, 193–205.
- Van Merriënboer, J. J. G., & Paas, F. (2003). Powerful learning and the many faces of instructional design: Toward a framework for the design of powerful learning environments. In E. De Corte, L. Verschaffel, N. Entwistle, & J. J. G. van Merriënboer (Eds.), *Powerful learning environments: Unravelling basic components and dimensions*. Oxford, UK: Elsevier Science.
- Vermunt, J. D. H. M. (1992). *Leerstijlen en sturen van leerprocessen in het hoger onderwijs: Naar procesgerichte instructie in zelfstandig denken* [Learning styles and regulations of learning in higher education: Toward process-oriented instruction in autonomous thinking]. Amsterdam/Lisse, The Netherlands: Swets & Zeitlinger.
- Vermunt, J. D. H. M. (1996). Metacognitive, cognitive, and affective aspects of learning styles and strategies: A phenomenographic analysis. *Higher Education*, 31, 25–50.
- Vermunt, J. D. H. M. (2003). The power of learning environments and the quality of student learning. In E. De Corte, L. Verschaffel, N. Entwistle, & J. J. G. van Merriënboer (Eds.), *Powerful learning environments: Unravelling basic components and dimensions*. Oxford, UK: Elsevier Science.
- Vermunt, J. D. H. M., Bouhuijs, P. A. J., & Picarelli, A. (2003). *Vragenlijst Leerstijlen voor het Voortgezet Onderwijs (VLS-VO)* [Inventory of learning styles for secondary education (ILS-SE)]. Maastricht, The Netherlands: Expertise Center Active Learning, Maastricht University.
- Vermunt, J. D. H. M., & Vermetten, Y. J. (2004). Patterns in student learning: Relationships between learning strategies, conceptions of learning, and learning orientations. *Educational Psychology Review*, 16, 359–384.
- Veugelers, W., de Jong, U., & Schellings, G. (2004). Studie naar het onderzoek van de tweede fase havo/vwo [Metastudy of research of the second phase in senior general secondary education/preuniversity education]. Amsterdam: Instituut voor de Lerarenopleiding.
- Weinstein, R. S. (1998). Promoting positive expectations in schooling. In N. M. Lambert & B. L. McCombs (Eds.), *How students learn: Reforming schools through learner-centered education* (pp. 81–111). Washington, DC: American Psychological Association.
- Wierstra, R. F. A., & Beerends, E. P. M. (1996). *Leeromgevingspercepties en leerstrategieën van eerstejaars studenten sociale wetenschappen* [Perceptions of the learning environment and learning strategies of social science students in their first year]. *Tijdschrift voor Onderwijsresearch*, 21, 306–322.
- Wierstra, R. F. A., Kanselaar, G., van der Linden, J. L., & Lodewijks, H. G. L. C. (1999). Learning environment perceptions of European university students. *Learning Environments Research*, 2, 79–98.
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, 25, 68–81.

Received August 29, 2006

Revision received December 18, 2007

Accepted January 18, 2008 ■

# Using Argumentation Vee Diagrams (AVDs) for Promoting Argument–Counterargument Integration in Reflective Writing

E. Michael Nussbaum  
University of Nevada, Las Vegas

This study examined a new prewriting tool, argumentation vee diagrams (AVDs), which are used to write reflective opinion essays. AVDs are based on the theoretical concept of argument–counterargument integration, which involves evaluating and integrating both sides of an issue before developing a final conclusion on a controversial question. In a test of the effectiveness of AVDs, 45 undergraduates at a large, southwestern university were randomly assigned to an experimental or control group. Both groups wrote 4 opinion essays over a 4-week period. The experimental group also received training on using the AVDs, including instruction on criteria for weighing arguments. Results indicated that AVD training was effective in enhancing argument–counterargument integration. Furthermore, examination of integration strategies used by participants revealed a new strategy, *minimization*, which was not previously part of E. M. Nussbaum and G. Schraw's (2007) argument–counterargument integration framework. Minimization involves curtailing the importance or extensiveness of a problem or advantage as a heuristic shortcut for weighing advantages and disadvantages. The role of critical questions and argumentation schemata in argument–counterargument integration is discussed.

**Keywords:** argumentation, critical thinking, graphic organizers, thinking skills, writing (composition)

The following is an example of argument–counterargument integration:

One argument for the U.S. invasion of Iraq was the elimination of weapons of mass destruction. Counterarguments included the possibility of destabilizing the region and the huge cost in dollars and human lives. On balance, the counterarguments now appear stronger. There was little evidence for the administration's claims, the costs have been tremendous, and more people than ever are joining terrorist groups. A possible "in-between" solution would have been to use a United Nations force so the United States would not appear to the Muslim world to be an aggressor.

Evaluating arguments for and against a course of action is an essential skill for effective personal and social decision making, as the above example illustrates. The ability to critically combine arguments and counterarguments into an overall final opinion is known as *argument–counterargument integration* (Nussbaum & Schraw, 2007; Nussbaum, Winsor, Aqui, & Poliquin, 2007). Argument–counterargument integration is an important aspect of writing reflective or persuasive essays. This study examined what strategies college students use when asked to engage in argument–counterargument integration. It also examined the effect of training students to engage in argument–counterargument integration using graphic organizers known as *argumentation vee diagrams* (AVDs).

---

E. Michael Nussbaum, Department of Educational Psychology, University of Nevada, Las Vegas.

I would like to thank Ordene V. Edwards and Jennifer D. Golanics for assistance in coding.

Correspondence concerning this article should be addressed to E. Michael Nussbaum, Department of Educational Psychology, Box 453003, University of Nevada, Las Vegas, 4505 Maryland Parkway, Las Vegas, NV 89154-3003. E-mail: nussbaum@unlv.nevada.edu

## Theoretical Foundations

### *Philosophic Foundations*

Several philosophic frameworks inform contemporary work on argumentation (Nussbaum, 2007). Philosophic models are intended to provide normative standards of good argumentation. Most contemporary models recognize that argumentative reasoning is both defeasible and dialectical. Defeasibility refers to the fact that a good argument is one that can withstand (i.e., defeat) objections (Pollock, 1987). Dialectical arguments are those that are constructed in response to questions or challenges generated by an interlocutor (argumentation can occur among individuals or as part of internal dialogue; see Kuhn, 2005). Philosophers disagree as to whether an argument must be generated dialectically to even qualify as an argument (Finocchiaro, 2005; Johnson, 2002), but most would agree that a good argument must respond to challenges (Erduran, Simon, & Osborne, 2004; Vorebej, 2006).

Naess (1966) proposed that one should analyze arguments by examining the reasons in support of a conclusion (pro-argumentation) and those supporting the opposite conclusion (contra-argumentation). Counterarguments can also show why a pro-argument is weak or flawed; such arguments have also been referred to as *refutations* (Aristotle, 350 B.C.E./1955), *undercutting defeaters* (Pollock, 1987), or *Counter-Cs* (Felton & Kuhn, 2001). Although there is not complete consistency in how various authors use these terms, Kuhn (1991) used the term *rebuttal* to refer to an argument refuting a counterargument, (i.e., a counter-argument to a counterargument; see also Nussbaum & Kardash, 2005). These are all Naess-like models.

Although not necessarily more valid, a more influential model of argument was advanced in the late 1950s by Toulmin (1958), who proposed that arguments made in ordinary discourse have six components: (a) *claim* (i.e., conclusion), (b) *grounds* (response to



the question, "Why?" or "What are your reasons?"), (c) *warrant* (response to the question, "So what?" or "How do your grounds lead to your claim?"), (d) *backing* ("How do you establish the truth of the warrant?"), (e) *rebuttals* (to potential counterarguments or exceptions), and (f) *modal qualifiers* (that qualify certainty in the conclusion). Toulmin initially proposed his model to show that formal logic provided an insufficient basis for judging the strength of arguments. It is unclear whether Toulmin intended his model primarily as a descriptive or normative framework; the model has been used in both ways (e.g., Kelly, Drunker, & Chen, 1998). Descriptively, it has been used as the basis for most argumentation coding systems in educational research (Bell & Linn, 2000; Erduran et al., 2004; Jiménez-Aleixandre, Rodríguez, & Duschl, 2000; Lizotte, McNeil, & Krajcik, 2004; Marttunen, 1994; Mason & Santi, 1994; Russell, 1983); prescriptively, it has been claimed that good arguments should contain Toulmin's various argumentation components (Chinn, 2006; Kenyon & Reiser, 2006; Knudson, 1994; Toth, Suthers, & Lesgold, 2002; Yeh, 1998). In the domain of writing, it has been used to encourage students to make their warrants more explicit and elaborate (Chambliss & Murphy, 2002; Rottenberg, 1988). Although explicitness is important, these efforts give short shrift to the consideration and integration of counterarguments (van Eemeren, Grootendorst, Jackson, & Jacobs, 1993; Willard, 1976). According to the Toulmin model, good arguments should contain a rebuttal to an implicit or briefly mentioned counterargument, but there is no emphasis on developing strong or multiple counterarguments.

Fortunately, more recent philosophic models of argumentation correct this problem. The pragma-dialectical school of argument (van Eemeren & Grootendorst, 1992) views arguments as constructed in dialectical exchanges known as *critical discussions*, the purpose of which is to resolve a difference of opinion in a rational manner. Another contemporary model of argumentation has been proposed by Douglas Walton (1995). In an analysis of a large selection of informal arguments, Walton (1996) identified different patterns of reasoning, which he called *argumentation schemes*. Associated with each pattern is a set of *critical questions* that a good reasoner should ask, such as "How strong is the likelihood that certain consequences will occur?" or "Are there consequences of the opposite value that should be taken into account?" Critical questions can also be found in the Toulmin model, but the Walton model contains a greater number and variety of critical questions, providing a stronger basis for pro- and contra-argumentation. Pro- and contra-argumentation in turn provides the basis for the theoretical concept explored in this article: argument-counterargument integration.

### *Argument-Counterargument Integration*

In the psychology of reasoning, an important consideration is how to make students' reasoning more balanced. Individuals often reason in biased ways, specifically by searching for evidence that supports prior beliefs and by ignoring counterevidence (Chinn & Brewer, 1998). Confirmation and belief biases are well documented in the literature on reasoning (Mayer, 1992). To overcome these biases, educators must teach students to reason in a balanced way, encouraging students to bracket and critique their biases (through asking critical questions) and to consider opposing views as much as their own views. For this reason, much attention has

been given in argumentation research to encouraging students to generate counterarguments (Kuhn, 2005; Nussbaum, 2005).

But it is important for students also to learn to critically evaluate arguments and counterarguments in order to reach an overall final conclusion. Nussbaum and Schraw (2007) termed this process *argument-counterargument integration*. Argument-counterargument integration is loosely based on neo-Piagetian views of reasoning development (Case, 1985; Halford & McCredde, 1998). This perspective views development as a process of coordinating and integrating disparate elements (in working memory) into a more coherent conceptual structure (McCutchen, 1996). In argumentation, the disparate elements are individual arguments and counterarguments that can then be integrated in various ways. Effective argumentation also involves metacognitive reflection, a "stepping back" that allows one to view and weigh the overall merits of different arguments and counterarguments (Kuhn, 1991, 2005), specifically by asking critical question (Walton, 1995). Argument-counterargument integration is also informed by (a) Suedfeld, Tetlock, and Streufert's (1992) work on measuring integrative complexity, which is the degree to which people make distinctions (*differentiation*) in discourse and conceptual connections among these distinctions (*integration*) and (b) the work of Leitão (2000), who examined replies to counterarguments in the context of everyday discourse and found examples of *integrative replies* that qualified original positions by acknowledging and integrating counterarguments.

### *Application to Writing*

Student writing assignments have long been viewed as a vehicle for promoting student learning and critical thinking (Knipper & Duggan, 2006; Pandis, Ward, & Matthews, 2005). Writing allows students to reflect on and draw linkages among ideas. Students also learn to formulate and critique written arguments.

*Persuasive versus reflective writing.* Most students are taught to formulate arguments during writing (if they are taught argumentation at all) in the context of *persuasive writing*. Students are frequently taught to develop a thesis and defend it by citing reasons and evidence in support of their position (Fulkerson, 1996). This is the basis of the infamous five-paragraph essay. Using less scripted techniques, educational researchers (e.g., De La Paz, 2005; Ferretti, MacArther, & Dowdy, 2000; Nussbaum & Kardash, 2005) have explored how to promote better writing among students of all ages by having them set goals and subgoals for their writing (for example, subgoals to provide reasons, evidence, explanation) and to engage in various self-regulatory activities to improve the persuasiveness of their essays (Graham, Harris, & Mason, 2005). Nevertheless, teaching persuasive writing tends to promote one-sided thinking. Students do not tend to consider counterarguments when writing persuasive essays (Santos & Santos, 1999; Stapleton, 2001), or they consider them only briefly.

Recently, there have been calls to teach students a different genre of argumentative writing, which I shall term *reflective writing* (see also Kim, Anderson, Nguyen-Jaheil, & Archodidou, 2007). Other names given to this different genre have included *deliberative writing* (Kroll, 2000), *analytic writing* (Nussbaum & Kardash, 2005), or *constructive argument* (Hewlett, 2006; Mallin



& Anderson, 2000). Reflective writing provides a context for argument-counterargument integration. The focus in reflective writing is on exploring and integrating various sides of an issue in order to reach a reasoned conclusion. In contrast to persuasive writing, in which the student starts with a conclusion and defends it, in reflective writing, the conclusion comes at the end (although it certainly can be summarized in the introduction). Reflective writing allows for a more in-depth exploration of a problem or question. It provides students with an opportunity to practice a greater variety of critical thinking operations.

*Collaborative argumentation.* Persuasive writing reflects an adversarial approach to argumentation in which one side attempts to win over another. Reflective writing, on the other hand, is more consistent with recent attention in argumentation research on collaborative reasoning (Anderson et al., 2001), which is said to promote a greater exploration of ideas (Keefer, Zeitz, & Resnick, 2000), more conceptual change (Chinn, 2006; Driver, Newton, & Osborne, 2000; Reznitskaya et al., 2001), and better problem-solving skills (Wegerif, Mercer, & Dawes, 1999). Much of this research has been conducted in the context of oral discussions, however, and not specifically in essay writing. Although essay writing is often an individual activity, it is still dialectical in the sense that the writer can pose questions to himself or herself and answer them, or otherwise have an internal dialogue (Vygotsky, 1978). Most philosophers now consider the construction of arguments an inherently dialectical activity (Freeman, 1992; Johnson, 2002), and from an educational psychology perspective, enhanced internal dialogue is related to better planning and self-regulation of the writing process (Harris, Graham, & Mason, 2006; Reznitskaya et al., 2001). Thus, the recent theoretical turn toward collaborative argumentation is also applicable to essay writing.

*Strategies for argument-counterargument integration.* How can the concept of argument-counterargument integration be applied to the writing of reflective essays? Nussbaum and Schraw (2007) identified several strategies that students can use to achieve argument-counterargument integration. First, in a weighing strategy, the arguer considers both sides and then considers which side has the stronger arguments (for example, Do advantages outweigh disadvantages?). Second, in a synthesis strategy, the arguer finds a final standpoint between different sides (e.g., finding a creative solution in which benefits are realized while disadvantages are minimized or in recognizing that the wisdom of an alternative may depend on certain factors; that is, it may be wise in some circumstances but not in others). The use of creative synthesis as an argumentation move is neglected in most current argumentation models (e.g., Toulmin, 1958; van Eemeren & Grootendorst, 1992), even though various educational psychologists have called for more attention to be given to creative thought (Bereiter & Scardamalia, 2006; Sternberg & Lubart, 1991).

A third strategy identified by Nussbaum and Schraw was *refutation*, showing that one or more arguments on a particular side of an issue are false, irrelevant, or insufficiently supported. Refutation is not a strong integration strategy because—while it addresses counterarguments—it tends to be associated with one-sided reasoning, rather than balanced reasoning. Nevertheless, refutation can be used in the context of balanced reasoning if used selectively (e.g., some arguments are refuted but others, on both sides, are weighed or synthesized).

### *Nussbaum & Schraw (2007) Study*

Most prior argumentation models have emphasized refutation over integration. As a result, there is a dearth of psychological research on how well individuals are able to integrate arguments and counterarguments. In an initial study, Nussbaum and Schraw (2007) asked 84 college undergraduates to write reflective essays on the topic, "Does watching TV cause children to become more violent?" The study had a  $2 \times 2$  factorial design, and two interventions were aimed at promoting argument-counterargument integration. The first was to briefly instruct students on the qualities of a good argument (i.e., arguments should have a clear position, adequate supporting reasons, compelling counterarguments with supporting reasons, and a final conclusion that reflects one of the argument-counterargument integration strategies). This instruction was provided in the context of an extended example, "Should Janet Jackson and Justine Timberlake be fined for the baring of Jackson's breast at the Superbowl?" The second intervention was to provide students with a graphic organizer to help plan their essays. The organizer consisted of several ovals for students to fill in: an argument oval (and ovals for supporting reasons), a counterargument oval (and ovals for supporting reasons), and an oval at the top for an overall final conclusion.

The researchers found that the integration strategies were enhanced when students were given instruction on the criteria of a good argument. On the other hand, use of a graphic organizer to plan essays only enhanced a refutation strategy. The organizer helped students manage working memory demands, making it easier for them to remember counterarguments and thus to rebut them.

It was not clear, however, why only refutation was enhanced. One possibility relates to how the graphic organizer was designed; the integration oval was placed at the top, showing that it was supported by the various arguments. Although students were instructed to complete the integration oval last, some may have completed it first, thus prematurely committing to a side. Another possibility was that students may have had more experience with the refutation strategy than the others (although data are needed to confirm this), so unless students were explicitly cued to use the other strategies, they tended not to do so. Furthermore, such cues may need to be integrated with the graphic organizer.

Alternatively, there may be something inherent about graphic organizers that facilitate refutation. Organizers make arguments-counterarguments salient, so it is easy to pick arguments to support and refute. The processes of weighing arguments, or constructing creative solutions, are harder to make visually salient. Furthermore, organizers invite students to think of several arguments on one side of an issue before those on the other side, and cognitive interference (from thoughts on one side) may make it harder to develop arguments on the other side as much, making it easier to refute. In the Nussbaum and Schraw (2007) study, students who received both interventions (criteria instruction and then the graphic organizer) still tended to use refutation over integration strategies, even though they received instruction in integration. An important question is whether the graphic organizer can be redesigned to facilitate integration strategies.



## The Present Study

The present study had three goals. First, given the failure of graphic organizers to promote integrative reasoning in the previous study, I sought to evaluate a longer term intervention in which students were given more help in understanding the purpose of the graphic organizer and more practice using it. This was done by redesigning the graphic organizer (described later), combining the organizer with more instruction specifically on how to integrate arguments and counterargument, and strengthening the instruction to include discussion of “critical questions” that should be asked when integration is performed. I hypothesized that the redesigned intervention would result in more balanced reasoning in the experimental group compared with that in a control group and that the effects would grow stronger over time.

The second goal was to understand the argumentation “moves” (or stratagems; Anderson et al., 2001) that individuals use when performing the integration task. This research question was open ended and intended to shed light on which moves participants naturally used (specifically, the moves by those in the control group) as well as those moves used after training. It was anticipated that some participants would use moves that were more normatively advanced than others. Identifying moves that participants found feasible to use can inform future training efforts. It was hypothesized that among participants who did engage in argument-counterargument integration, the most common strategy would be synthesis (developing creative solutions), because synthesis creates a more unified mental representation that poses lower cognitive load than weighing, which requires maintenance of a number of disparate elements in working memory. Weighing was anticipated to be the strategy used the least.

The final goal was to better understand participants’ instructional histories in regard to writing opinion essays. One explanation of the Nussbaum and Schraw (2007) finding of a tendency to use biased reasoning with graphic organizers was that biased reasoning was the default mode used, specifically because that is the type of persuasive reasoning typically taught in schools. However, no data were collected on participants’ instructional histories, so in the present study, a survey was included in this regard. It was hypothesized that some students would have had some prior training in refuting arguments but little or no training in other, more balanced integration strategies (i.e., weighing costs and benefits, or designing a creative solution).

## Method

### Participants

Participants were 45 undergraduates at a large, southwestern university recruited from educational psychology courses. Most were education majors (53% elementary, 18% secondary, and 14% other); the remaining were majoring in other disciplines (13%) or were undecided (2%). Almost all (89%) were seeking a teaching credential and were either juniors (56%) or sophomores (38%). The average grade point average was 3.16. The ethnicity breakdown was 66% European American, 16% Hispanic/Latino, 9% Asian/Asian American, 5% African American, and 4% other. In respect to gender, 84.4% were women, and 15.6% were men. Students participated in the experiment to meet a course requirement.

## Materials

**Writing prompts.** There were four writing prompts related to (a) requiring school uniforms, (b) accountability systems, (c) ability grouping, and (d) grading for effort or participation. Each prompt mentioned some arguments that students could use (pro and con), but gave no guidance on integration except for instructing students to write arguments on both sides and then develop and support their opinion. The complete prompts are contained in Appendix A. To randomize the order that the prompts were given to the participants, I randomly assigned each participant to one of four different prompt sequences; the participants received one prompt per session. The prompt orders were as follows:

Order 1: uniforms, accountability, grading, ability grouping.

Order 2: ability grouping, grading, accountability, uniforms.

Order 3: grading, uniforms, ability grouping, accountability.

Order 4: accountability, ability grouping, uniforms, grading.

**Graphic organizer (AVD).** Figure 1 shows the redesigned organizer. The form was called an *argumentation vee diagram* (AVD) and was adapted from a form originally used by Novak and Gowin (1984) to structure science investigations. Unlike the form used in Nussbaum and Schraw (2007), the integration section of the form in the present study was placed at the bottom, so that students would be less tempted to fill it out first. In addition, two critical questions were included on the form to cue various integration strategies: “Which side is stronger, and why?” and “Is there a compromise or creative solution?” Students could also use a refutation strategy by drawing an arrow between an argument and the counterargument that refuted it (or a rebuttal that refuted the counterargument) and were informed of this possible response both in the training and in the written instructions.

To determine whether placing one of the critical questions first made any difference, we developed a second version of the AVD form (i.e., Form B). With this form, the first critical question asked was, “Is there a compromise or creative solution?” followed by the question, “Which side is stronger, and why?” We randomly assigned students to use one of the two versions of the form and, as will be reported later, the version of the form that the students used had no discernable effect.

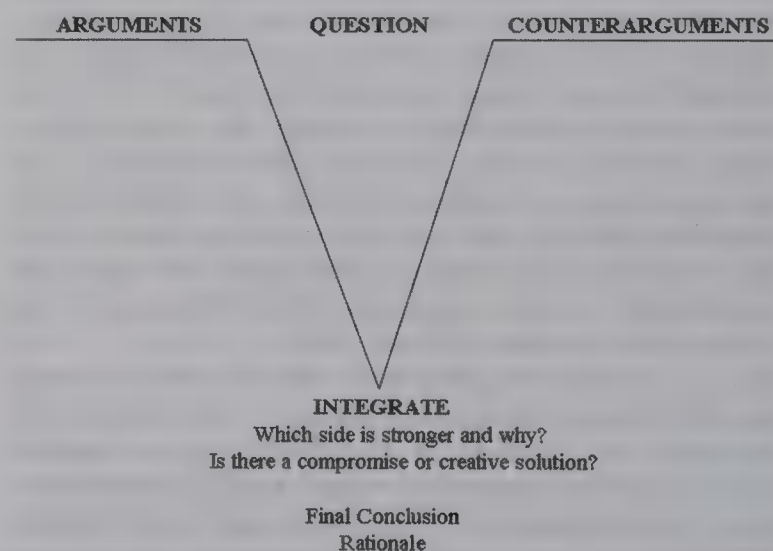


Figure 1. Argumentation vee diagram. The question is written in the vee. Arguments and counterarguments are paraphrased in the spaces outside the vee. A final conclusion and argument is written at the base of the vee.

*Training materials.* For training purposes, overheads (and handouts) of a completed AVD were prepared on a question related to whether the university should lease parking spaces to faculty and students (see Figure 2). A whiteboard was used to complete an AVD on the topic of whether watching TV causes children to become more violent, with a blank AVD projected onto the whiteboard.

*AVD evaluation form.* Students in the experimental group also completed a survey after using the AVDs regarding whether they perceived the AVDs as helpful and whether participants were given sufficient time to complete them. The evaluation, which consisted of six open-ended questions, is shown in Appendix B. It took about 3–4 min to complete.

*Prior experience or instruction with integration.* We administered an 18-item survey to assess the students' prior experience and instruction with the different integration strategies. Students rated, on a 4-point Likert scale, their prior experience with such things as discussing counterarguments or designing creative solutions. Appendix C lists the exact items. The survey took about 5 min to complete.

### Design and Procedure

This study was to some extent a design experiment (Brown, 1992; Design-Based Research Collective, 2003), in which small changes were made to the intervention in mid-course to improve its effectiveness. These changes will be described later. In addition,

a repeated measures design was used, with participants randomly assigned to either the experimental ( $n = 22$ ) or control group ( $n = 23$ ). Participants attended four 1-hr sessions; during each session, students wrote for 30 min on one of the four topics (the order was selected randomly for each student). The amount of remaining time was announced after 15 min and 25 min. I conducted all the sessions. During Session 1, the purpose of the study was explained, as follows:

This study examines opinion essays, where one examines both sides of an issue and then takes a stand. The purpose of the study is to examine how people write opinion essays and the effects of repeated practice. You will write one essay per session and may also receive some feedback and training.

Participants then completed informed consent forms and then the demographics survey. They were then told that they would have 30 min to write an essay; this essay served as a baseline measure. They were also told that everyone would be writing on a different topic. After removing their prompt from a manila envelope (which contained all their materials), they were given the following instructions:

Your essay should give reasons for and against each viewpoint and should then present and justify your viewpoint. Don't worry about spelling and grammar. You may spend a few minutes planning what you will write. If you complete your essay early, please sit quietly until everyone is done.

The procedures for the experimental and control groups were the same up to this point. The session for the control group was then concluded. The experimental group, on the other hand, was introduced to AVDs via a worked example on leasing university parking spaces (see Figure 2). The integration section of the example illustrated both a weighing and a synthesis strategy. An overhead projector was used to show students examples of both the blank and completed AVDs shown in Figures 1 and 2. Then, participants were asked as a group to give arguments on the other training topic (whether TV causes violence) and—as a group—constructed a completed AVD on the whiteboard.

During Session 2, the parking AVD was reviewed with the experimental group; they received copies of both a blank AVD and a completed one. Experimental group students were asked to individually complete a practice AVD on TV violence. They were also reminded that it was not necessary to answer both integration questions listed on the AVD, just one or the other. Participants were then given 8 min to complete an AVD on their writing topic for that week and 30 min to prepare their essay. The period of 8 min was judged as sufficient on the basis of a pilot session, but a question on this issue was asked on the AVD evaluation surveys, which were completed at the end of the session by the experimental group.

During Session 3, students received written feedback on the two AVDs that they had prepared the previous week. Also, the weighing strategy was discussed in more detail; students were asked, in the context of the TV violence training question, "What sort of things generally make one side stronger than another?" The list generated from the question was discussed. I then mentioned that *prevalence* (the extent of an advantage or disadvantage) and *value* (the importance of a specific advantage or disadvantage) should be relevant considerations, along with evidence. I also told the stu-

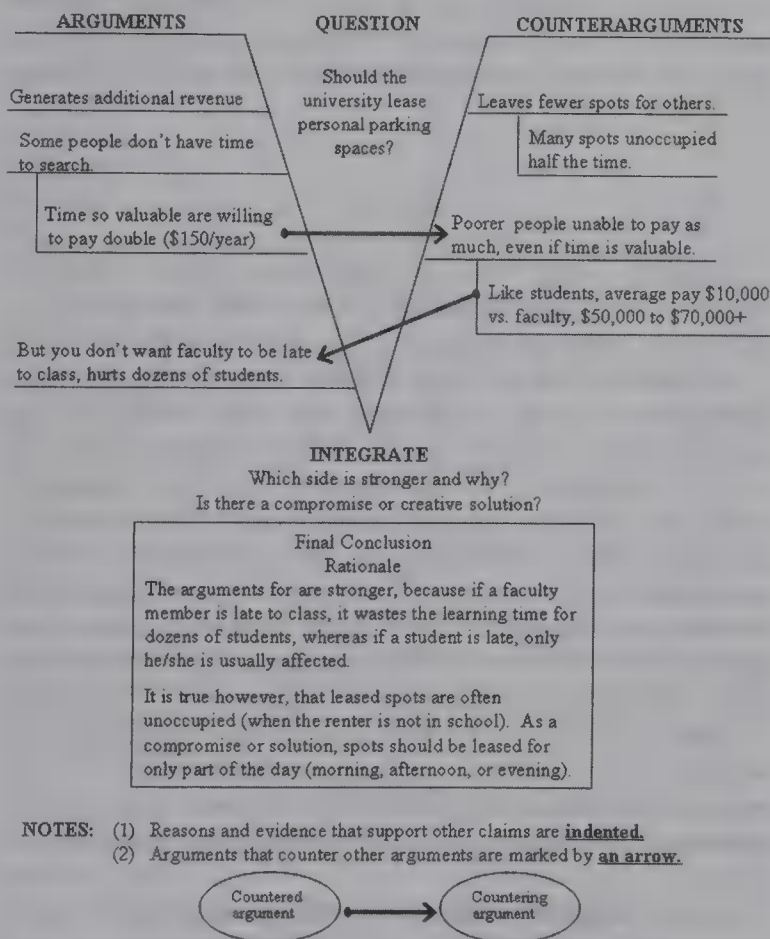


Figure 2. Completed argumentation vee diagram on the question, "Should the university lease personal parking spaces to faculty and students?"



dents that the arguments did not have to be completely pro or con to be put on the AVD but that they had to decide on which side each argument would best fit. Students then prepared their essays, planning them with the AVDs as before. They also completed the prior instruction/feedback survey and another AVD evaluation form.

The fourth session was used to assess internalization and transfer of the integration strategies; students wrote on a new topic based on the prompt assigned to them that week, but they were not provided with AVDs. They were given 8 min to plan their essay.

Unlike the experimental group, the control group participants were not given any instruction in argument-counterargument integration or in essay writing. As with the experimental group, the control group was told that the purpose of the study was to see whether the quality of their arguments improved with practice. Like the experimental group, they were given 8 min of planning time before writing each essay but were not provided with AVDs. They were instructed—both orally and as part of their writing prompts—to address both sides of an issue and then to take a stand, giving their opinion along with supporting reasons.

The control group did not complete the prior experience and instruction form as this could have cued them to use integration strategies. One goal of the study was to examine what strategies students naturally use.

### Coding

Each essay was coded for the presence of one of the three integration strategies: specifically weighing (contrasting the relative merits of an argument or counterargument), synthesis, and refutation. Each category was coded dichotomously, indicating the presence or absence of that strategy. The coding system allowed for the possibility that a student could use more than one strategy. After I had performed some preliminary coding, it became clear that an additional category called *pseudo-integration* was needed. In pseudo-integration, students indicated that the case for a particular side was strongest but supported their assertion simply by picking a supporting argument they liked best. This is not true integration, because the integration ignores how to address counterarguments. The pseudo-integration category was subdivided into either just restating, at the end of an essay, the strongest argument versus adding onto the argument (*amplifying*) to make it stronger (e.g., adding examples and explanations, or explaining why a consequence is important). Students who used one of the three strategies (synthesis, weighing, or refutation) received a 0 for pseudo-integration, since it was conceptualized as an alternative to integration.

Appendix D contains excerpts from essays reflecting the different coding categories. Brief summaries of the examples are given below. All scoring was performed blind to condition and session.

**Synthesis.** Synthesis was indicated by either (a) a claim describing how an alternative should be designed to alleviate negative consequences (e.g., “Accountability systems should use multiple types of assessment, so the curriculum will not be overly narrow”), or (b) contingent claims (e.g., “It depends”) that specify the conditions under which the arguments apply and the conditions under which the counterarguments apply (e.g., “Whether ability grouping is a good educational practice would depend on whether students can move up to a higher group”).

**Weighing.** A weighing strategy was indicated by either (a) an argument that a value associated with one side is more important than a value on the other side (e.g., “In grading, rewarding effort is more important than measuring K–12 content knowledge because, in the long run, it is effort that will lead to career success”), (b) an argument that a certain advantage is more prevalent than a disadvantage or vice versa or that a certain advantage or disadvantage is not all that widespread (e.g., on the school uniforms topic: “Conflicts related to clothing are infrequent”), or (c) an argument that an advantage on one side could be realized by the other side (e.g., “In grading, class participation promotes learning as much as just assessing knowledge because students are learning social skills”).

**Refutation.** In refutation, the writer makes an argument for why a counterargument is weak or flawed (e.g., in writing on the school uniforms topic, a writer refutes the counterargument that freedom of expression is suppressed by noting that students can express their persona in other ways, such as through hair style). Refutation removes the “force” of the counterargument by showing that it not that applicable or likely or that it rests on a faulty assumption.

**Pseudo-integration.** In pseudo-integration, a writer presents both sides of an issue and then (typically in the final paragraph) indicates which side he or she agrees with and why. However, no reference to the arguments on the other side is made. The writer just restates or amplifies arguments made on the favored side (see Appendix D for examples).

**Interrater reliability.** All essays were independently scored by two raters: the researcher and a graduate assistant who was unaware of the study’s hypotheses. All disagreements were resolved through discussion. Interrater agreement for all scoring categories was 80% or higher.

### Results

Frequency counts of the strategies used are shown in Table 1. There were five cases of students who used more than one strategy in an essay; these are indicated in the footnotes to the table.

For purposes of the statistical analysis, scores for synthesis were aggregated with those for weighing and called *integration*. Because the critical questions on the AVDs prompted students for these two strategies, it was hypothesized that overall integration associated with these strategies would increase. This hypothesis did not differentiate between the two strategies (the use of one or both could increase), so to test the hypothesis, I combined scores from the two categories. (There was a separate hypothesis that synthesis would be used more than comparison by both groups, but this conjecture was not tied to the treatment and so will be discussed later.) The use of the term *integration* for synthesis and weighing is not meant to imply that refutation does not also involve some form of integration, but it is a weak form (counterarguments are acknowledged and evaluated, but ultimately rejected).

Because almost all variables were dichotomous, logistic regression was used to analyze the results. Logistic regression is more valid than analysis of variance (ANOVA) when the outcome variable is dichotomous because of nonnormality and nonuniform variances; also, linear methods assume the outcome variable can

Table 1  
*Integration Strategies Used: Frequency by Session and Condition*

Strategy	Group	
	Control	Experimental
Session 1		
None	8	6
Pseudo-integration		
Restatement	2	5
Amplification	3	3
Refutation	0	3
Integration <sup>a</sup>		
Synthesis	8	5
Weighing	2	2
Subtotal	23	24 <sup>b</sup>
N	23	22
Session 2		
None	10	3
Pseudo-integration		
Restatement	3	3
Amplification	3	7
Refutation	1	3
Integration <sup>a</sup>		
Synthesis	4	5
Weighing	2	1
Subtotal	23	22
N	23	22
Session 3		
None	9	2
Pseudo-integration		
Restatement	0	1
Amplification	2	3
Refutation	4	2
Integration <sup>a</sup>		
Synthesis	5	11
Weighing	1	5 <sup>c</sup>
Subtotal	21	24
N	21	21
Session 4		
None	6	4
Pseudo-integration		
Restatement	2	3
Amplification	2	2
Refutation	3	2
Integration <sup>a</sup>		
Synthesis	6	3
Weighing	1	4
Subtotal	20	18
N	20	18

<sup>a</sup> Not including refutation. <sup>b</sup> Reflects 1 participant who used both a synthesis and weighing strategy, and 1 participant who used both a refutation and weighing strategy. <sup>c</sup> Reflects 2 participants who used both a synthesis and weighing strategy, and 1 participant who used both a refutation and weighing strategy.

keep increasing beyond the value of 1 and so are misspecified (Agresti, 2002).

The integration variable in this study ranged from 0 to 2; however, a 2 was assigned in only three cases. Because the presence of 2 was so infrequent, I simplified the analysis by making the variable dichotomous (replacing the 2 with a 1), so that the variable merely reflected the presence or absence of integration.

The lack of independence among observations from the same person was controlled with hierarchical linear modeling (HLM) and MLwiN statistical software (Rabash, Steele, Browne, & Prosser, 2005); Time was used as Level 1 and Participant as Level 2 (i.e., Time was nested in Participant). HLM is increasingly being used to analyze time series data (Tan, 2008) and can be applied to binary outcomes (Rabash et al., 2005).

### Results of Intervention

We first consider the results of the first three sessions, because the fourth session was a transfer session in which no AVDs were used. Only data for the 42 participants who attended all three sessions were used. There was not a significant difference in the frequency of the prompt orders between the two conditions,  $\chi^2(3, N = 42) = 0.98, p = .81$ , indicating the attempt to balance the prompt orders was successful.

As shown in Figure 3 and Table 2, the effects of the intervention over time were positive. With use of integration strategies as the dependent measure, there was a Time (Session)  $\times$  Condition interaction,  $B = 1.31, p < .01$ , odds ratio (OR) = 3.71, with a significant difference between the experimental and control groups for the third session favoring the experimental group,  $\chi^2(1, N = 38) = 4.68, p < .05$ , Cramer's  $V = .38$ . Specifically, 67% of the essays in the experimental group were integrated, while only 29% of the control essays were. There was not a significant Time  $\times$  Treatment interaction for refutations,  $B = -2.29, p = ns$ , OR = .10, or pseudo-integration,  $B = -0.03, p = ns$ , OR = 0.97.

These effects did not occur during the transfer session (Session 4), when the AVDs were not used. There was no significant difference for integration,  $\chi^2(1, N = 42) = 0.02, p = ns$ , Cramer's  $V = .02$ .

There was no evidence that the amount of integration on the AVD-planned essays was affected by the order of the integration questions (putting "Which side is stronger, and why?" vs. "Is there a compromise or creative solution?" first). Half the participants ( $n = 11$ ) received one type of AVD, and the other half ( $n = 11$ ) received the AVD with the question order switched. Participants who received one type of AVD during Session 2 received the other type during Session 3. Results of Mann-Whitney exact tests showed the order of questions had no statistical effect within the experimental group on integration during either Session 2 ( $U = 57.5, p = .85$ ) or Session 3 ( $U = 50.0, p = .76$ ).

### Nature of Integration Strategies

In order to better understand the nature of the integration strategies that were used, I conducted further analysis by subtype. Integration strategies can involve synthesizing a creative solution or weighing (i.e., weighing advantages against disadvantages).

Table 3 presents the means and standard deviations of the strategies used. Table 4 presents the counts of the different integration strategies aggregated over the four sessions, along with the proportion that each strategy was used. The total figures show that about one half of the students engaged in either no integration (27.4%) or pseudo-integration (25.1%). Of those who engaged in integration, synthesis (e.g., finding a creative solution) was the most common strategy (26.9%), as expected. Synthesis possibly creates an integrated representation that places a lower burden on



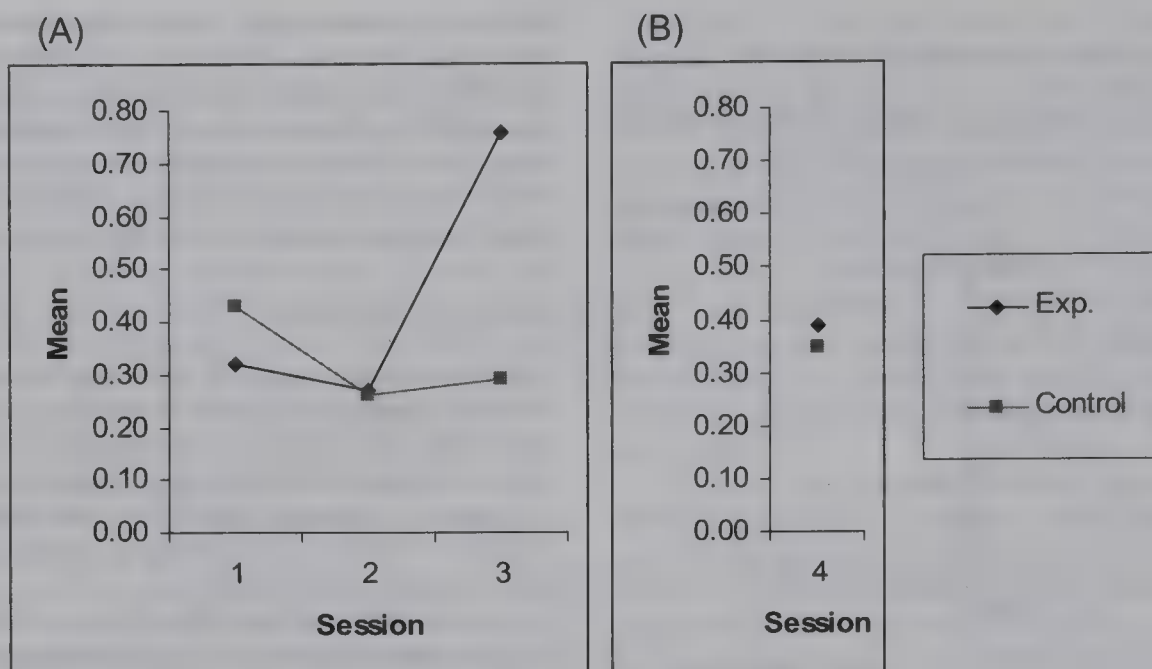


Figure 3. Mean use of integration strategies, comparing experimental (Exp., diamonds) and control (squares) groups for Sessions 1–3 (A) and for the transfer session, in which no AVDs were used by the experimental group (B).

working memory than do the other strategies. As expected, weighing was uncommon (10.3%) but was tied with refutations (10.3%). Because a prior hypothesis was made about the relative frequency of synthesis versus weighing, the difference was subjected to a statistical test and found to be significant,  $\chi^2(1, N = 42) = 12.94$ ,  $p \leq .001$ .

Because these figures could have been affected by the intervention, I also examined figures for just the control group, and a similar picture still emerged. Synthesis and pseudo-integration were again the most frequently used strategies (26.4% and 19.5%, respectively). With pseudo-integration, it was common for students to pick what they considered to be the strongest argument and to elaborate on it further (providing additional reasons, examples, or explanation of importance), a process that has been termed *amplification* (see Perelman & Olbrechts-Tyteca, 1958/1969). Amplification occurred 43% more often compared with just re-statement of the strongest argument. In the control group, the least

used strategy was weighing,  $\chi^2(1, N = 21) = 9.97$ ,  $p \leq .002$ , compared with synthesis.

Although students rarely used weighing strategies, this does not mean that such strategies are unimportant; these strategies can be quite effective in identifying the best course of action, as implied by classical decision theory (Baron, 1988). Nevertheless, such strategies must also be psychologically realistic, meaning that they can be used by individuals with some amount of ease. Analysis of how participants actually used weighing strategies in this experiment can shed light on this issue. Weighing was not comprehensive; that is, students—at least explicitly—did not consider the total weight of arguments on one side and balance them against those on the other. Instead, they reasoned more locally, where just one argument was contrasted with another on the other side. (The strategy would be considered a heuristic shortcut, which facilitates judgment while minimizing cognitive demands; see Goldstein & Gigerenzer, 2002.) In the total corpus, weighing strategies were used 18 times. In 53% of these cases, students mentioned two arguments (on different sides of an issue) and indicated that one argument reflected a more important value. In another 10% of the cases, the students made an argument that a value was realized on both sides of an issue. In the remaining 38% of cases, the students argued that a particular disadvantage was not that widespread or important (for example, “Children’s free expression is not that important; school is about fitting in”) and contrasting it with another value (for example, “avoiding being teased about wearing cheap clothes”), a process that was termed *minimization* (in contrast to amplification). This was not an integration strategy that had been anticipated; however, we did discuss problem prevalence during Session 3 (and 83% of the minimization cases occurred in the experimental group during or after Session 3). I therefore tested whether, in the experimental group, there

Table 2  
Logistic Model Parameters for Integration Strategies Used

Source	B	OR <sup>a</sup>	p
Condition	−1.83	0.16	.08
Time <sup>b</sup>	0.50	1.65	.17
Time × Condition	1.31**	3.71	.008

Note. Parameters for multilevel logistic model ( $N = 42$ , with three observations per participant). Level 1 reflected time; Level 2 reflected participant.

<sup>a</sup> OR = odds ratio (odds of using an integration strategy associated with one of the predictors set to 1 divided by the odds with predictor set to 0). For Time, OR reflects ratio from a one unit increase in T. OR is estimated by exponentiating the B coefficient ( $e^B$ ). <sup>b</sup> Time = Session 1, 2, or 3.

\*\*  $p < .01$ .

Table 3  
Strategies Used: Means (and Standard Deviations)

Strategy	Control group		Experimental group	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Overall				
None	0.38	0.49	0.18	0.39
Pseudo-integration				
Restatement	0.08	0.27	0.14	0.35
Amplification	0.11	0.32	0.18	0.39
Refutation	0.09	0.29	0.12	0.33
Integration <sup>a</sup>	0.33	0.47	0.44	0.57
Session 1				
None	0.39	0.49	0.27	0.46
Pseudo-integration				
Restatement	0.09	0.29	0.23	0.43
Amplification	0.13	0.34	0.14	0.35
Refutation	0.00	0.00	0.14	0.35
Integration <sup>a</sup>	0.43	0.51	0.32	0.57
Session 2				
None	0.43	0.51	0.14	0.35
Pseudo-integration				
Restatement	0.13	0.34	0.14	0.35
Amplification	0.13	0.34	0.32	0.48
Refutation	0.04	0.21	0.14	0.35
Integration <sup>a</sup>	0.26	0.45	0.27	0.46
Session 3				
None	0.43	0.51	0.10	0.31
Pseudo-integration				
Restatement	0.00	0.00	0.05	0.21
Amplification	0.10	0.31	0.14	0.36
Refutation	0.19	0.40	0.10	0.30
Integration <sup>a</sup>	0.29	0.46	0.76	0.62
Session 4				
None	0.30	0.47	0.22	0.43
Pseudo-integration				
Restatement	0.10	0.31	0.17	0.38
Amplification	0.10	0.31	0.11	0.32
Refutation	0.15	0.37	0.11	0.32
Integration <sup>a</sup>	0.35	0.50	0.39	0.50

<sup>a</sup> Not including refutation.

was a statistically significant trend for minimization to occur more frequently in the latter sessions; the test was significant ( $\tau = 0.22, p \leq .01$ ). This is evidence that use of this stratagem may have been affected by the instruction.

### Prior Experience and Instruction

Table 5 shows the means and standard deviations for the prior experience and instruction survey items. The table shows that participants had a moderate amount of experience (*M*s close to 2.0) in discussing and rebutting counterarguments and in writing persuasive essays that did not discuss counterarguments or that discussed them only briefly. Thus, as expected, writing persuasive essays was a familiar writing genre for these students. However, so was writing essays in which two or more sides of an issue were explored in a balanced manner ( $M = 1.78$ ) and in which arguments and counterarguments were weighed ( $M = 1.83$ ). So participants

may have had some experience in reflective writing. However, the finding that they had experience weighing arguments against counterarguments was surprising because participants did not generally use explicit weighing strategies when writing their essays. However, pseudo-integration (just picking the argument that seems strongest and elaborating on it) was used the most frequently, and participants may have interpreted weighing in this fashion. It is interesting they had the least amount of experience and instruction with the two synthesis strategies (designing creative solutions or compromises between two sides), with means of 1.39 and 1.22, respectively, for experience and means of 1.22 and 1.33, respectively, for instruction. These figures reflect the response indicating *a little* experience/instruction (1.0) in synthesis and were statistically different from the response indicating *moderate* experience/instruction (2.0) on the basis of one-sample *t* tests,  $t(17) = -2.83$  for experience in creative design,  $t(17) = -4.08$  for instruction in creative design,  $t(17) = -3.76$  for experience in developing compromises,  $t(17) = -3.37$  for instruction in developing compromises,  $p \leq .01$  for all tests. Thus, the hypothesis that participants had little experience with and instruction in integration strategies was for the most part verified, particularly in relationship to synthesis.

### Participant Evaluations of AVDs

Although all participants in the experimental group were asked to fill out the AVD evaluation form, a few failed to do so because of time constraints. Usable responses were  $n = 21$  for Session 2 and  $n = 19$  for Session 3.

In response to Item 1, 95% of the participants indicated that they were given sufficient time to complete their AVDs during Session 2, but this figure dropped to 68% during Session 3 (percentage giving an unqualified "yes"). The remainder desired a few more minutes to think about the integration, but some noted that they were allowed to use some of the 30 min devoted to essay writing to complete their AVD. The decrease, which was statistically significant ( $p < .05$ , Fisher's exact test), may indicate that participants were responding to the weighing criterion that was introduced during Session 3; one participant noted that "we had more things to think about."

Table 4  
Strategies Used: Frequency and Percentage

Strategy	Control group		Experimental group		Total	
	<i>f</i>	%	<i>f</i>	%	<i>f</i>	%
None	33	37.9	15	17.0	48	27.4
Pseudo-integration:	17	19.5	27	30.7	44	25.1
Restatement	(7)	(8.0)	(12)	(13.6)	(19)	(10.9)
Amplification	(10)	(11.5)	(15)	(17.0)	(25)	(14.3)
Refutation	8	9.2	10	11.4	18	10.3
Integration <sup>a</sup>	29	33.3	36	40.9	65	37.1
Synthesis	(23)	(26.4)	(24)	(27.3)	(47)	(26.9)
Weighing	(6)	(6.9)	(12)	(13.6)	(18)	(10.3)
Total	87	100.0	88	100.0	175	100.0

Note. Parentheses indicate that values are subsets of the value for the overall strategy.

<sup>a</sup> Not including refutation.



Table 5  
*Prior Experience and Instruction in Key Argumentation Tasks*

Item	Experience <sup>a</sup>		Instruction <sup>b</sup>	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Discussing counterarguments (reasons others might disagree with you)	1.94	0.80	1.50	0.79
Rebutting counterarguments (reasons the counterarguments are wrong)	1.72	0.89	1.39	0.85
Weighing arguments and counterarguments (to decide which side is stronger)	1.83	0.79	1.67	0.84
Designing solutions that address both arguments and counterarguments	1.39	0.92	1.22	0.80
Writing essays that are designed to persuade others of a point of view but that do not briefly discuss counterarguments	1.83	0.86	1.50	0.79
Writing essays that are designed to persuade others of a point of view and that do briefly discusses counterarguments	2.06	0.73	1.61	0.92
Writing essays that explores two or more sides of an issue in a balanced manner (with about the same amount of discussion of each side).	1.78	1.00	1.50	0.92
Designing a compromise between two sides.	1.22	0.88	1.33	0.84
How much experience have you had in writing opinion essays?	2.11	0.76	—	—
How much instruction have you had in writing opinion essays?	—	—	1.89	0.76

*Note.* *N* = 18 (given to experimental group only). Response options: *None* = 0, *a little* = 1, *moderate* = 2, and *a lot* = 3.  
<sup>a</sup> Stem: “Before this study, please rate how much experience you have had doing each of the following while writing opinion essays.” <sup>b</sup> Stem: “Before this study, please rate how much instruction you have received regarding each of the following in relationship to writing opinion essays.”

All participants indicated that they were given sufficient time to prepare their essays (Item 2). In regards to Item 3, most thought—without qualification—that the oral and written directions for how to complete the AVDs were clear (78% for Session 2, 80% for Session 3). The remainder had various minor questions that were answered, at least for Session 2, during the following session; there was no common theme to any of these questions.

Item 4 asked what participants found beneficial about using argumentation vee diagrams. The responses were grouped inductively into categories; results are shown in Table 6. The dominant response was that the AVDs assisted students in organizing their essays. A few participants were more specific and indicated that the AVDs helped them “separate the two sides” or “keep track of which counterarguments cancelled out which arguments.” On the remaining evaluation items, no participant made any points repeated by other participants, so no themes were identified.

Discussion

This article presents argument–counterargument integration as a worthwhile task for students to engage in when writing reflective opinion essays. The study found that graphic organizers (argumentation vee diagrams) and instruction in strategies for integrating

arguments and counterarguments may have initially promoted only pseudo-integration, where students, in concluding their essays, picked one side of the issue and a supporting argument that they felt was strongest. This was not true integration because counterarguments, although previously stated, were then ignored. True integration, however, did emerge over time, with additional practice and instruction. The effect disappeared when support from the argumentation vee diagrams was withdrawn. Future research is needed to determine whether transfer effects can be obtained with a longer (e.g., semester-long) intervention. It is encouraging that there was still more pseudo-integration in the experimental group than in the control group during the transfer session, as this might signal the beginnings of some independent skill acquisition.

The hypothesis that writing reflective (rather than persuasive) essays would be a new genre of writing for these students was not confirmed; they reported some experience writing essays exploring two sides of an issue and even weighing advantages and disadvantages. However, an ability to explicitly explain why one side of an issue was stronger than another was not apparent in most student essays, so it is likely that such weighing just reflects pseudo-integration. In addition, students reported that they had received very little instruction in writing essays that propose compromises or creative solutions between two sides or instruction in how to develop refutations. Therefore, it appears that students could benefit from specific instruction in argument–counterargument integration.

In what follows, I first discuss specific conclusions relating to how this instruction should be designed. I then discuss more general issues regarding the theoretical and practical significance of the findings.

Design Issues

*Strength standards.* This study was to some extent a design experiment (Brown, 1992; Design-Based Research Collective, 2003), in which small changes were made to the intervention mid-course to improve its effectiveness. Specifically, after the first session, it was clear that students needed more instruction in standards of what makes one argument stronger than another, so

Table 6  
*Participant Comments on Benefits of Argumentation Vee Diagrams*

Category	Percentage	
	Session 2 <sup>a</sup>	Session 3 <sup>b</sup>
Assists with organization	36.4	73.7
Assists with brainstorming	13.6	15.8
Assists with time management	9.1	0.0
Helps in remembering and “seeing” thoughts	22.7	10.6
Helps in separating/correcting the sides	18.2	5.3
Helps in seeing which side is stronger	4.5	10.6
Unclear response	9.1	5.3

<sup>a</sup> *n* = 21. <sup>b</sup> *n* = 19.

they could avoid pseudo-integration. Because most of the discussion questions involved practical reasoning—that is, deciding among different courses of action (Walton, 1990)—classical decision theory was used as a guide.

According to classical decision theory (von Neumann & Morgenstern, 1944), individuals should choose the alternative that maximizes the expected utility for either themselves (according to classical economic models, Smith, 1980) or the community at large (according to social utilitarian models, Bentham, 1879/2000). Expected utility is a product of both the likelihood that various positive and negative consequences will occur from a particular course of action and the value attached to those consequences. Although there are strong debates in psychology as to what extent individuals can reason probabilistically (Anderson, 1990; Griffiths & Tenenbaum, 2006; Tversky & Kahneman, 1983), individuals can likely make qualitative judgments as to whether particular consequences are likely or unlikely, as well as valuable or less valuable (Grennan, 1997). Therefore, students were instructed to consider the prevalence and importance of particular consequences. Prevalence refers to how widespread a particular consequence would be (i.e., frequency); it is an easier way psychologically for individuals to think about probabilities (Goldstein & Gigerenzer, 2002).

*Graphic organizer design.* The graphic organizer used in this study was a redesign of one tested in a previous study (Nussbaum & Schraw, 2007) that found that the organizer resulted in more refutational (but not integrative) reasoning. That was not the case with the present study, indicating that the changes made to the graphic organizer (moving the integration section to the bottom of the form and adding integrative questions) were effective in improving the design. Nevertheless, the instruction that was provided to students could also have accounted for this effect. The focus in design research, however, is to design the strongest intervention possible, on the assumption that multiple components are necessary (interactively) to produce an effective intervention (Design-Based Research Collective, 2003). Additional cycles of redesign and implementation will be needed, however, to determine which aspects of the intervention are the most critical in producing integrative reasoning. It seems likely that AVDs used in the absence of any accompanying instruction would produce only pseudo-integration, given the prevalence of pseudo-integration in this study even in the experimental condition.

### Theoretical Significance

This section situates the theoretical significance of the findings in both philosophic and psychological models of argumentation.

*Philosophic models.* As noted in the introduction, Walton (1996) identified 22 patterns of reasoning, which he called *argumentation schemes*. Associated with each pattern is a set of *critical questions* that a good reasoner should ask. For example, one common scheme, and the one most applicable to the reasoning questions asked in this study, is *argument from consequences*, in which a course of action is supported by citing good consequences that will result from it or is rejected because of the bad consequences. Associated with *arguing from consequences* are three critical questions, specifically (from Walton, 1996, pp. 76–77): (a) “How strong is the likelihood that these cited consequences will occur?” (b) “What evidence supports the claim that the conse-

quences will occur?” and (c) “Are there consequences of the opposite value that should be taken into account?”

Walton’s critical questions provide some criteria for evaluating arguments (Nussbaum, 2007). Arguments that answer these questions are considered stronger because they take more considerations into account. The research reported in this article is most consistent with the Walton model but extends the model in that two additional critical questions are posed. The first is, “Can a creative solution be designed to realize the advantages of one side while minimizing the disadvantages of the other?” A number of educational psychologists have called for more attention to be paid to design thinking (Bereiter & Scardamalia, 2006; Sternberg & Lubart, 1991); these authors, as well as the results of this study, have suggested that it is rarely taught in schools. Another critical question suggested by the present study is, “How prevalent are the consequences of a particular course of action?” (lack of prevalence is the strategy of minimization). This is really a restatement of Walton’s first critical question but one that may be easier for students to master because it is one that was actually used by several reasoners in this study.

The Walton model, as amended here, may be a useful thing to teach to students. Specifically, students can learn to evaluate arguments by learning the key, critical questions that they should ask for different patterns of reasoning. Although many teachers attempt to teach critical thinking and critical writing, they lack well-defined instructional models of what factors constitute these skills (Paul, Elder, & Bartell, 1997). The framework elaborated in this article can provide the basis of critical writing instruction based in argumentation.

*Psychological models.* The initial theoretical basis of this study was neo-Piagetian theory, which contends that the development of reasoning consists of the coordination and integration of disparate elements in working memory (Case, 1985; Halford & McCredde, 1998). The development of functional working memory capacity and the creation of central conceptual structures (Case, 1992) have explained developmental phenomena in mathematics, literacy, and social and spatial reasoning (Case & Okamoto, 1996). In addition, knowledge integration has been proposed as a central goal of scientific reasoning (Linn & Hsi, 2000). So it seemed sensible to propose the integration of arguments and counterarguments as a worthwhile goal as well. Nevertheless, as this study proceeded, it became clear that students needed criteria with which to judge the strength of arguments.

In this study, Walton’s philosophic model became the basis of these criteria, suggesting that schema theory may be a useful supplement to the idea of argument-counterargument integration. Walton’s schemes are also psychological schemata that can be inculcated in students. The schemata have slots for different components of arguments (e.g., arguments, counterarguments, rebuttals) but also critical questions that should be asked about different components (essentially psychological operations associated with different slots). These schemata must be developed in students as a goal of instruction and can function as integrating devices.

While more research is needed to support this account, it is consistent with the work of Anderson and colleagues (Reznitskaya & Anderson, 2002), who identified a number of argumentation moves, called *stratagem*, that were used by students with increasing frequency in literature discussions (Anderson et al., 2001). Reznitskaya and Anderson (2002) claimed that stratagems are the



building blocks of argumentation schemata, which they used to explain transfer effects from oral discussion to writing. Stratagems are the psychological (and social) operations associated with different schemata. The present study is theoretically significant in identifying specific stratagem, such as minimization, that can usefully be taught to students, specifically as critical questions. Students can learn to ask themselves how widespread certain benefits or harms are or whether a creative solution can be designed to mitigate those harms. The present study shows that the concept of argument-counterargument integration likely needs to be supplemented with notions of schema in order to be theoretically and pedagogically powerful.

### Conclusion

As noted previously, much work on argumentation in education is based on the Toulmin model. Although this has yielded some informative research studies, it may have also constrained theoretical development in the field of argumentation research somewhat (Nussbaum, 2007). Use of alternative frameworks, such as those provided by argument-counterargument integration and schema theory, can provide new theoretical insights. The purpose of this research was to begin to develop (empirically and theoretically) an alternative framework. A corresponding limitation is that the framework is in an early stage of development and so needs more elaboration. Another limitation of this research is that the sample size for the survey on prior instruction was small, and so future research studies should seek to replicate these findings on a larger sample.

Future studies should also be conducted over a longer time frame. In this study, the progress made by students in the experimental group was not maintained when the AVDs were not used in Session 4. Longer term studies are therefore needed to determine how much instructional time is required for the argumentation schemas reflected in the AVDs to be internalized by learners and used spontaneously. Longer term studies should also identify whether students can be explicitly taught a minimization strategy. That was a useful argumentation stratagem that occurred in the study, but students were not taught it explicitly and it was not used on a wide scale. Finally, future research studies might attempt to disentangle the relative contributions of the graphic organizer from the verbal explanation, examples, and feedback that were provided with it. The effect of providing students with feedback on the extent and quality of their integration (but without their using AVDs) should also be explored.

A final limitation of the study was that prior to scoring, I did read the AVDs of the experimental group for Session 2 for the purpose of providing feedback, and although this was done several weeks before scoring, it is possible that some memory of these arguments may have provided cues as to which essays were in the experimental group, possibly biasing the scoring. It should be noted that no significant difference between the experimental and control groups was found for Session 2, however, so any unintentional biasing would have made no substantive difference in the results.

This study shows that complex critical thinking can be taught to students. A focus on key critical questions that students should ask themselves during writing, while integrating various sides of an issue, can help build the necessary schemata to promote more

reflective writing and ultimately, we hope, better decision making. Effective decision making is the cornerstone of successful democratic governance, and as the opening quote to this article on the Iraq war illustrates, it can also often be a matter of life and death.

### References

- Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley-Interscience.
- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Anderson, R. C., Nguyen-Jahiel, K., McNurlen, B., Archodidou, A., Kim, S., Reznitskaya, A., et al. (2001). The snowball phenomenon: Spread of ways of talking and ways of thinking across groups of children. *Cognition and Instruction*, 19, 1–46.
- Aristotle (1955). *On sophistical refutations* (E. S. Forster & D. J. Furley, Trans.). Cambridge, MA: Harvard University Press. (Original work published 350 B.C.E.)
- Baron, J. (1988). *Thinking and deciding*. New York: Cambridge University Press.
- Bell, P., & Linn, M. C. (2000). Scientific arguments as learning artifacts: Designs for learning from the Web with KIE. *International Journal of Science Education*, 22, 797–817.
- Bentham, J. (2000). *An inquiry into the nature and causes of the wealth of nations*. Oxford, England: Clarendon Press.
- Bereiter, C., & Scardamalia, M. S. (2006). Education for the knowledge age. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 695–714). Mahwah, NJ: Erlbaum.
- Brown, A. L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences*, 2, 141–178.
- Case, R. (1985). *Intellectual development: Birth to adulthood*. New York: Academic.
- Case, R. (1992). *The mind's staircase: Exploring the conceptual underpinnings of children's thought and knowledge*. Hillsdale, NJ: Erlbaum.
- Case, R., & Okamoto, Y. (1996). The role of central conceptual structure in the development of children's thought. *Monographs of the Society for Research in Child Development*, 61, (1–2, Serial No. 246).
- Chambliss, M. J., & Murphy, P. K. (2002). Fourth and fifth graders representing the argument structure in written texts. *Discourse Processes*, 34, 91–115.
- Chinn, C. A. (2006). Learning to argue. In A. M. O'Donnell, C. E. Hmelo-Silver, & G. Erkens (Eds.), *Collaborative learning, reasoning, and technology* (pp. 355–383). Mahwah, NJ: Erlbaum.
- Chinn, C. A., & Brewer, W. F. (1998). An empirical test of a taxonomy of responses to anomalous data in science. *Journal of Research in Science Teaching*, 35, 623–654.
- De La Paz, S. (2005). Effects of historical reasoning instruction and writing strategy mastery in culturally and academically diverse middle school classrooms. *Journal of Educational Psychology*, 97, 139–156.
- Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher*, 32, 5–8.
- Driver, R., Newton, P., & Osborne, J. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education*, 84, 287–312.
- Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science Education*, 88, 915–933.
- Felton, M., & Kuhn, D. (2001). The development of argumentative discourse skill. *Discourse Processes*, 32, 135–153.
- Ferretti, R. P., MacArthur, C. A., & Dowdy, N. S. (2000). The effects of an elaborated goal on the persuasive writing of students with learning disabilities and their normal achieving peers. *Journal of Educational Psychology*, 92, 694–702.



- Finocchiaro, M. A. (2005). *Arguments about arguments: Systematic, critical and historical essays in logical theory*. New York: Cambridge University Press.
- Freeman, J. B. (1992). *Dialectics and the macrostructure of arguments*. New York: Foris.
- Fulkerson, R. (1996). *Teaching the argument in writing*. Urbana, IL: National Council of Teachers of English.
- Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, 109, 75–90.
- Graham, S., Harris, K. R., & Mason, L. (2005). Improving the writing performance, knowledge, and self-efficacy of struggling young writers. *Contemporary Educational Psychology*, 30, 207–241.
- Grennan, W. (1997). *Informal logic: Issues and techniques*. Montreal, Quebec, Canada: McGill-Queen's University Press.
- Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Research*, 17, 767–773.
- Halford, G. S., & McCredden, J. E. (1998). Cognitive science questions for cognitive development: The concepts of learning, analogy, and capacity. *Learning and Instruction*, 8, 289–308.
- Harris, K. R., Graham, S., & Mason, L. H. (2006). Improving the writing, knowledge, and motivation of struggling young writers: Effects of self-regulated strategy development with and without peer support. *American Educational Research Journal*, 43, 295–340.
- Hewlett, A. K. (2006). Constructive thinking from theory to practice: An exploratory study (Doctoral dissertation, University of Manitoba). *Dissertation Abstracts International*, 66, 3551.
- Jiménez-Aleixandre, M. P., Rodríguez, A. B., & Duschl, R. A. (2000). "Doing the lesson" or "doing science": Argument in high school genetics. *Science Education*, 84, 757–792.
- Johnson, R. H. (2002). *Manifest rationality: A pragmatic theory of argument*. Mahwah, NJ: Erlbaum.
- Keefer, M. W., Zeitz, C. M., & Resnick, L. B. (2000). Judging the quality of peer-led dialogues. *Cognition and Instruction*, 18, 53–81.
- Kelly, G. J., Druker, S., & Chen, C. (1998). Students' reasoning about electricity: Combining performance assessments with argumentation analysis. *International Journal of Science Education*, 20, 849–871.
- Kenyon, L., & Reiser, B. J. (2006, April). *A functional approach to nature of science: Using epistemological understandings to construct and evaluate evidence*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.
- Kim, I.-H., Anderson, R. C., Nguyen-Jahiel, K., & Archodidou, A. (2007). Discourse patterns during children's collaborative online discussions. *Journal of the Learning Sciences*, 16, 333–370.
- Knipper, K. J., & Duggan, T. J. (2006). Writing to learn across the curriculum: Tools for comprehension in content area classes. *Reading Teacher*, 59, 462–470.
- Knudson, R. E. (1994). An analysis of persuasive discourse: Learning how to take a stand. *Discourse Processes*, 18, 211–230.
- Kroll, B. M. (2000). Broadening the repertoire: Alternatives to the argumentative edge. *Composition Studies*, 28, 11–27.
- Kuhn, D. (1991). *The skills of argument*. Cambridge, England: Cambridge University Press.
- Kuhn, D. (2005). *Education for thinking*. Cambridge, MA: Harvard University Press.
- Leitão, S. (2000). The potential of argument in knowledge building. *Human Development*, 43, 323–360.
- Linn, M. C., & Hsi, S. (2000). *Computers, teachers, peers: Science learning partners*. Mahwah, NJ: Erlbaum.
- Lizotte, D. J., McNeil, K. L., & Krajcik, J. (2004). Teacher practices that support students' construction of scientific explanations in middle school classrooms. In Y. Kafai, W. Sandoval, N. Enyedy, A. Nixon, & F. Herrera (Eds.), *Proceedings of the sixth international conference of the learning sciences* (pp. 310–317). Mahwah, NJ: Erlbaum.
- Mallin, I., & Anderson, K. V. (2000). Inviting constructive argument. *Argumentation and Advocacy*, 36, 120–133.
- Marttunen, M. (1994). Assessing argumentation skills among Finnish university students. *Learning and Instruction*, 4, 175–191.
- Mason, L., & Santi, M. (1994, April). *Argumentation structure and meta-cognition in constructing shared knowledge at school*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Mayer, R. E. (1992). *Thinking, problem solving, cognition* (2nd ed.). New York: Freeman.
- McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review*, 8, 299–325.
- Naess, A. (1966). *Communication and argument: Elements of an applied semantics*. London: Allen and Unwin.
- Novak, J. D., & Gowin, D. B. (1984). *Learning how to learn*. New York: Cambridge University Press.
- Nussbaum, E. M. (2007). *Beyond Toulmin: Argumentation frameworks for analysis and practice*. Las Vegas: University of Nevada.
- Nussbaum, E. M., & Kardash, C. M. (2005). The effects of goal instructions and text on the generation of counterarguments during writing. *Journal of Educational Psychology*, 97, 157–169.
- Nussbaum, E. M., & Schraw, G. (2007). Promoting argument-counterargument integration in students' writing. *Journal of Experimental Education*, 76, 59–92.
- Nussbaum, E. M., Winsor, D. L., Aqui, Y. M., & Poliquin, A. M. (2007). Putting the pieces together: Online argumentation vee diagrams enhance thinking during discussions. *International Journal of Computer-Supported Collaborative Learning*, 2, 479–500.
- Nussbaum, E. M. (2005). The effect of goal instructions and need for cognition on interactive argumentation. *Contemporary Educational Psychology*, 30, 286–313.
- Pandis, M., Ward, A., & Matthews, S. R. (Eds.). (2005). *Reading, writing, thinking: Proceedings of the 13th European conference on reading*. Newark, DE: International Reading Association.
- Paul, R. W., Elder, L., & Bartell, T. (1997). *California teacher preparation for instruction in critical thinking: Research findings and policy recommendations*. Sacramento, CA: California Commission on Teacher Credentialing. (ERIC Document No. 437 379)
- Perelman, C., & Olbrechts-Tyteca, L. (1969). *The new rhetoric: A treatise on argumentation* (J. Wilkinson & P. Weaver, Trans.). Notre Dame, IN: University of Notre Dame Press. (Original work published 1958)
- Pollock, J. L. (1987). Defeasible reasoning. *Cognitive Science*, 11, 481–518.
- Rabash, J., Steele, F., Browne, W., & Prosser, B. (2005). *A user's guide to MLwiN* (Version 2.0). London, England: Centre for Multilevel Modelling, University of Bristol.
- Reznitskaya, A., & Anderson, R. C. (2002). The argumentation schema and learning to reason. In C. C. Block and M. Pressley (Eds.), *Comprehension instruction: Research-based best practices* (pp. 319–334). New York: Guilford.
- Reznitskaya, A., Anderson, R. C., McNurlen, B., Ngyuen-Jahiel, K., Archodidou, A., & Kim, S. (2001). Influence of oral discussion on written argument. *Discourse Processes*, 32, 155–175.
- Rottenberg, A. T. (1988). *Elements of argument*. New York: St. Martin Press.
- Russell, T. L. (1983). Analyzing arguments in science classroom discourse: Can teachers' questions distort scientific authority? *Journal of Research in Science Teaching*, 20, 27–45.
- Santos, C. M. M., & Santos, S. L. (1999). Good argument: Content and contextual dimensions. In G. Rijlaarsdam & E. Espéret (Series Eds.) & J. Andriessen & P. Coirier (Vol. Eds.), *Foundations of argumentative text processing* (pp. 75–95). Amsterdam: Amsterdam University Press.
- Smith, A. (1980). *An inquiry into the nature and causes of the wealth of nations*. Chicago: Encyclopedia Britannica.



- Stapleton, P. (2001). Assessing critical thinking in the writing of Japanese university students. *Written Communication*, 18, 506–548.
- Sternberg, R. J., & Lubart, T. I. (1991). Creating creative minds. *Phi Delta Kappa*, 72, 608–615.
- Suedfeld, P., Tetlock, P. E., & Streufert, S. (1992). Conceptual/integrative complexity. In C. P. Smith (Ed.), in association with J. W. Atkinson, D. C. McClelland, & J. Veroff, *Motivation and personality: Handbook of thematic content analysis* (pp. 393–400). New York: Cambridge University Press.
- Tan, F. E. S. (2008). Best practices in analysis of longitudinal data: A multilevel approach. In J. E. Osborne (Ed.), *Best practices in quantitative methods* (pp. 451–470). Thousand Oaks, CA: Sage.
- Toth, E. E., Suthers, D. D., & Lesgold, A. M. (2002). “Mapping to know”: The effects of representational guidance and reflective assessment on scientific inquiry. *Science Education*, 86, 264–286.
- Toulmin, S. (1958). *The uses of argument*. New York: Cambridge University Press.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293–325.
- van Eemeren, F. H., & Grootendorst, R. (1992). *Argumentation, communication, and fallacies*. Hillsdale, NJ: Erlbaum.
- van Eemeren, F. H., Grootendorst, R., Jackson, S., & Jacobs, S. (1993). *Reconstructing argumentative discourse*. Tuscaloosa, AL: University of Alabama Press.
- von Neumann, J., & Morgenstern, O. (1944). *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press.
- Vorebej, M. (2006). *A theory of argument*. New York: Cambridge University Press.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Trans. & Eds.). Cambridge, MA: Harvard University Press.
- Walton, D. E. (1990). *Practical reasoning: Goal-driven, knowledge-based, action-guiding argumentation*. Totowa, NJ: Rowman & Littlefield.
- Walton, D. N. (1995). *A pragmatic theory of fallacy*. Tuscaloosa, AL: University of Alabama Press.
- Walton, D. N. (1996). *Argument schemes for presumptive reasoning*. Mahwah, NJ: Erlbaum.
- Wegerif, R., Mercer, N., & Dawes, L. (1999). From social interaction to individual reasoning: An empirical investigation of a possible socio-cultural model of cognitive development. *Learning and Instruction*, 9, 493–516.
- Willard, C. A. (1976). On the utility of descriptive diagrams for the analysis and criticism of arguments. *Communication Monographs*, 43, 309–319.
- Yeh, S. S. (1998). Empowering education: Teaching argumentative writing to cultural minority middle school students. *Research in the Teaching of English*, 33, 49–83.

## Appendix A

### Writing Prompts

#### Accountability

The Federal government mandates that every state have an accountability system by which schools are held accountable for how their students perform. For example, schools may be given greater or fewer funds on the basis of overall student performance on standardized tests. Also, some students in “underperforming” schools may be allowed to transfer. Advocates argue that accountability systems give schools an incentive to improve, may encourage or require more services and options to be provided to at-risk students, and provide parents and policy makers with information on year-to-year growth. Critics argue that accountability systems tend to narrow the curriculum, may punish schools that need the most help (in cases where funding is reduced), and may use indicators that are not totally valid.

In your opinion, should states be required to have an accountability system by which schools are held accountable for their students’ performance?

#### School Uniforms

Some people argue that public school students should be required to wear uniforms to school. Mandatory school uniform proponents argue that clothing is often a source of conflict in school, perhaps inciting theft and gang violence and also maintaining or widening the gap between those who can afford more expensive wardrobes and those who cannot. Requiring students to wear uniforms, they argue, not only removes this source of conflict but engenders a healthy attitude toward authority and

may make students take their education more seriously. Opponents of mandatory school uniform policies argue that school uniforms do not effectively deal with socioeconomic or cultural conflicts associated with clothing and that uniform policies discourage individuality and suppress freedom of expression.

In your opinion, should students in public schools be required to wear school uniforms? Think of as many reasons as you can on both sides of the issue when developing your opinion.

#### Ability Grouping

A third-grade teacher, Mr. Garcia, is planning to group his students into three reading groups (i.e., high reading ability, moderate reading ability, and low reading ability) on the basis of the results of a standardized reading test given to the students at the end of second grade. Is this good educational practice? Advocates of same-level ability grouping argue that it allows teachers to use materials and go at a pace appropriate for each student. Opponents of same-level ability grouping claim that it prevents the better students from acting as models for the weaker students.

What do you think? Think of as many reasons as you can on both sides of the issue when developing your opinion.

#### Grading

Is it good educational practice to base a portion of a student’s grade on such things as class participation, effort, or completion of

homework? Advocates of this practice claim that doing so gives students an incentive to try. Opponents argue that grades should only reflect how much a student has learned.

What do you think? Think of as many reasons as you can on both sides of the issue when developing your opinion.

### Additional Text

[Added to the end of each prompt was the following:] You will have 30 min to write an essay that discusses the issue and presents your opinion. (Do not worry about spelling and grammar.)

## Appendix B

### AVD Evaluation Form

Please answer each of the following questions regarding your experience today using an argumentation vee diagram (AVD).

1. Were you given sufficient time to complete the argumentation vee diagram? If not, please explain.

2. Were you given sufficient time to write your essay? If not, please explain.

3. Were the oral and written directions of how to complete an argumentation vee diagram clear? Are there any specific portions that you feel need more explanation? Please be as

specific as possible in regards to what you don't understand, if anything.

4. What do you find beneficial about using argumentation vee diagrams?

5. Is there anything you found unhelpful about using argumentation vee diagrams?

6. Do you have any specific suggestions about how we might improve the diagram or the procedure for using them? (You may write on the reverse side of this form.)

## Appendix C

### Experience and Instruction Survey Items

[The survey items were scored on a 4-point Likert scale, with response options: *none*, *a little*, *moderate*, and *a lot*.]

1. Before this study, please rate how much experience you have doing each of the following while writing opinion essays:

a. Discussing counterarguments (reasons others might disagree with you).

b. Rebutting counterarguments (reasons the counterarguments are wrong).

c. Weighing arguments and counterarguments (to decide which side is stronger).

d. Designing solutions that address both arguments and counterarguments.

e. Writing an essay designed to persuade others of a point of view but which does not discuss counterarguments?

f. Writing an essay designed to persuade others of a point of view and which briefly discusses counterarguments.

g. Writing an essay that explores two or more sides of an issue in a balanced manner (with about the same amount of discussion of each side).

h. Designing a compromise between two sides.

i. How much experience have you had in writing opinion essays?

2. Before this study, please rate how much *instruction* you have received regarding each of the following in relationship to writing opinion essays: [same items as in Question 1, except Item E was as follows: How much instruction have you previously received in writing opinion essays?]

## Appendix D

### Coding Examples

Below are more complete examples of each coding category, excerpted from the essays analyzed in the study. Author's comments are in brackets.

#### Integration Strategies

##### Synthesis

*Creative solution.* Concluding paragraph: I feel that states should have an accountability system to measure educational performance, but not have it be just off standardized test. Testing is not the best way to measure individual learning. Along with

testing, schools should consider how many students participate in extracurricular activities that the school offers. Like sports and clubs. They should also view classroom participation and social interactions. Schools need to be a catalyst for making the students better overall people in society. Some students may not test well, or the test may have biases. Testing is a good form to get an idea of a group's knowledge, but when looking at a school's overall performance, you must look at other factors as well. It's important to see how the school is involved and intertwined within the community. (Participant 15)



*Contingent ("It depends").* I feel that ability grouping is an excellent idea. It's good educational practice for the students as well as for the teachers. I also feel that it all depends on how the teachers want to organize and structure the ability group, so that the students would not get bored or feel uncomfortable about their reading skills. I also feel that the students should feel some type of achievement of their reading skills in knowing that they can improve and move on to the next reading level. The teacher should make the students feel engaged with their reading groups as well as feel and be engaged to improve their reading skills. (Participant 13)

### *Weighing*

*One value more important than another.* [In discussing whether students should be graded on class participation, effort, and completion of homework, the participant argues that effort is more important than knowledge]. . . Is it fair that my peers get the same grade as I do without "the effort?" I think the real question here is, what is the goal of education? If learning the facts is the main goal, then yes, it is fair that we get the same grade for acquiring the same amount of knowledge. Does our grade indicate our success in the future? I think not. I believe that my hard work and dedication will pay off in the end and that eventually the grades I got won't matter.

If the goal of education is simply to teach for the purpose of knowledge than the rest is not important. But if the purpose of education is to have successful professionals in society than yes, it is very important to give credit where credit is deserved. Some of us may not be math geniuses, but we try as hard as we can and hope that it is enough. If educators put more emphasis on effort, then the knowledge will eventually come. (Participant 23)

*Minimization.* [In comparing alternatives, a minimizing argument is one that shows that the advantages or disadvantages of one course of action will be limited.]

Clothing, I believe, has some to do with conflicts in school. Students group themselves with others who act, dress, and speak similarly to themselves. Clothing may blur the lines of different socioeconomic status, but only for a brief moment [minimizes the extent of the benefit]. Therefore, uniforms may not change much. Uniforms may, however, help students focus more in class by getting rid of distractions. Distractions such as logos and proper fit might influence some students. Uniforms could also create a more professional and conservative atmosphere for students to grow up in. I think that school uniforms is a good idea, but I don't think it is such a detrimental aspect to the students' educational success [minimizes the harm of clothing]. It is possible that uniforms could suppress student individuality and influence student rebellion, but not to an extreme [minimizes harm]. Mandatory school uniforms in public schools is a good idea but not a necessity. (Participant 46)

### *Refutation*

Even though students are prohibited the freedom to wear what they choose and maybe suppressed to "express" their speech through dress style, they are still able to express themselves based on how they carry out their persona. Their personal dress style may still be shown on how a student wears their hair, make up, and shoes. By no means is individuality being suppressed because their

uniqueness is greater seen through their actions rather than dress styles. (Participant 12)

## *Pseudo-Integration*

### *Restatement*

[At the conclusion, the writer picks one side and justifies it by restating arguments that were already made. In the following example, paragraphs are numbered for ease of reference.]

1. Should teachers and schools be held accountable by the Federal government for "underperforming" students in schools? Advocates argue that this accountability system gives schools incentives to improve. They also argue that it provides struggling students with more services. Critics argue that this accountability system narrows the curriculum and punishes schools that need the most help. Teachers and schools should not be held accountable for these underperforming schools due to the negative effects that they have on the students.

2. Advocates for accountability systems believe that the system will cause schools to improve. If a school is underperforming, the government cuts funds to the school to make the school want to improve. This is the opposite of what the government should do. The government should give more money to the "at-risk" schools, and spend more time training and preparing the teachers at the underperforming school rather than firing them. Advocates also argue that the system keeps the public up-to-date on how every school is doing. If a school does bad one year, then parents will pull their kids out of that school and will put them in a "smart" school, which will also cause many problems.

3. Accountability systems cause a narrowing of the curriculum at many schools. Many schools cut subjects such as science and social studies to focus more on language arts and math which is on the test. This limits the teacher on teaching to their full capability and it limits the students on crucial knowledge that they need in the future. Schools that narrow the curriculum have teachers that "teach to the tests." Teachers only teach what they know will be on the test, and that is not fair for the students.

4. Accountability systems should not shut down underperforming schools. The government should get help for these schools and provide them with the things that they need to help them improve in the future. Shutting down schools hurts a community, and mostly the students in the schools. Students who failed at an underperforming school will most likely fail at an adequate school also. In this situation, no one wins.

5. [All ideas in this concluding paragraph were made before. The opposing side is ignored.] Teachers and schools should not have accountability systems. These systems limit teachers, administrators, and many students [summary of Paragraph 3]. "Underperforming" schools should receive help from the government, not punishment [summary of Paragraph 4]. Teachers should not have to narrow their curriculum just so their students can pass the standardized tests [point made in Paragraph 3]. Accountability systems are narrowing the future of many struggling schools and students in America. (Participant 37)

### *Amplification*

[In this category, the writer, in the conclusion, picks one of the sides and elaborates on at least one of the supporting arguments to

make it more compelling and convincing. In the following example, the amplifications in the concluding paragraph are shown in italics.]

Some believe that a portion of a student's overall grade should account class participation effort and/or completion of home work while others would disagree to this as good educational practice. Primarily effort and overall knowledge reflection are the opposing concerns.

Proponents for such grading criteria suggest doing so gives students incentives to try. If such requirements aren't expected, the students would have little to work towards. Considering an overall grade influenced by such acts as participating in class encourages involvement which may potentially heighten a learning experience. In addition, students may see a completed homework assignment rewarding and will have a sense of accomplishment with a final product.

Conversely, opponents view these not as perks but rather downfalls to an overall knowledge development. Requiring such assignments may cause anxiety to get a good grade on homework with the knowledge that doing poorly can affect their grades. This anxiety may be a total disruption hindering the student from concentration. Proponents also believe that grades should only be

determined by reflecting on what a student has learned. If such mandates were determined, concepts may be taught to the homework as opposed to an overall learning, narrowing the scope of each subject. Also, students might begin to compare work with one another, building a competitive attitude.

. . . In my opinion, I believe participation, effort, and homework completion should influence an overall grade. *Students are at a young age they need goals set for them and incentives monitoring progress. Some students are by nature reclusive; with the notion of participation, students may open up more than they have otherwise. Giving expectations of students can push them to achieve better.* It is a proper way to monitor and reflect on our teaching, providing useful information. Educational practice considering these factors, and not based solely on these requirements, can be beneficial for both students and educators. [Italic emphasis added, indicating elaboration and additional points in first paragraph relating to incentives.]

Received June 11, 2007

Revision received February 5, 2008

Accepted February 17, 2008 ■

## Low Publication Prices for APA Members and Affiliates

**Keeping you up-to-date.** All APA Fellows, Members, Associates, and Student Affiliates receive—as part of their annual dues—subscriptions to the *American Psychologist* and *APA Monitor*. High School Teacher and International Affiliates receive subscriptions to the *APA Monitor*, and they may subscribe to the *American Psychologist* at a significantly reduced rate. In addition, all Members and Student Affiliates are eligible for savings of up to 60% (plus a journal credit) on all other APA journals, as well as significant discounts on subscriptions from cooperating societies and publishers (e.g., the American Association for Counseling and Development, Academic Press, and Human Sciences Press).

**Essential resources.** APA members and affiliates receive special rates for purchases of APA books, including the *Publication Manual of the American Psychological Association*, and on dozens of new topical books each year.

**Other benefits of membership.** Membership in APA also provides eligibility for competitive insurance plans, continuing education programs, reduced APA convention fees, and specialty divisions.

**More information.** Write to American Psychological Association, Membership Services, 750 First Street, NE, Washington, DC 20002-4242.



# Predictors of Word Decoding and Reading Fluency Across Languages Varying in Orthographic Consistency

George K. Georgiou and Rauno Parrila  
University of Alberta

Timothy C. Papadopoulos  
University of Cyprus

Very few studies have directly compared reading acquisition across different orthographies. The authors examined the concurrent and longitudinal predictors of word decoding and reading fluency in children learning to read in an orthographically inconsistent language (English) and in an orthographically consistent language (Greek). One hundred ten English-speaking children and 70 Greek-speaking children attending Grade 1 were examined in measures of phonological awareness, phonological memory, rapid naming speed, orthographic processing, word decoding, and reading fluency. The same children were reassessed on word decoding and reading fluency measures when they were in Grade 2. The results of structural equation modeling indicated that both phonological and orthographic processing contributed uniquely to reading ability in Grades 1 and 2. However, the importance of these predictors was different in the two languages, particularly with respect to their effect on word decoding. The authors argue that the orthography that children are learning to read is an important factor that needs to be taken into account when models of reading development are being generalized across languages.

**Keywords:** reading development, longitudinal studies, cross-linguistic, psycholinguistic grain size theory

A wealth of evidence has established the prominent role of phonological processing in reading acquisition. Three different aspects of phonological processing—phonological awareness, phonological short-term memory, and rapid automatized naming (RAN)—predict the rate of reading acquisition in several alphabetic languages varying in orthographic consistency (e.g., de Jong & van der Leij, 1999; Holopainen, Ahonen, & Lyytinen, 2001; Muter, Hulme, Snowling, & Stevenson, 2004; Parrila, Kirby, & McQuarrie, 2004; Wagner & Torgesen, 1987). Researchers have tended to assume that the models of early reading development generalize across languages (e.g., Frith, 1985; Marsh, Friedman, Welch, & Desberg, 1981) despite the lack of cross-linguistic studies. In addition, little is known about the role of orthographic processing, defined as the ability to use visual-orthographic information in processing words in early reading development (e.g., Barker, Torgesen, & Wagner, 1992; Burt, 2006). We suggest that recent theoretical developments, the psycholinguistic grain size theory (PGST) and the theories of how RAN is related to reading, lead to expectations that both phonological processing and orthographic processing skills contribute differently to reading development in languages varying in orthographic consistency. In what follows, we first review the PGST, the RAN theories, and the predictions derived from them. We then summarize results of studies on the effects of phonological and orthographic processing

skills on reading across languages. Finally, we present an overview of the current study.

## The Psycholinguistic Grain Size Theory of Reading Development

Recently, Ziegler and Goswami (2005) introduced the psycholinguistic grain size theory (PGST), according to which “the dramatic differences in reading accuracy and reading speed found across orthographies reflect fundamental differences in the nature of the phonological recoding and reading strategies that are developing in response to the orthography” (p. 19). On one hand, children who are learning to read in orthographically consistent languages, such as Finnish, Greek, German, or Italian, rely heavily on grapheme–phoneme recoding strategies because the relationship between graphemes and phonemes is straightforward. On the other hand, children learning to read in orthographically inconsistent languages, such as English or Danish, cannot rely on smaller grain sizes because inconsistency is much higher for smaller grapheme units than for larger units. The reduced reliability of small grain sizes leads children to develop flexible unit size recoding strategies, such as grapheme–phoneme correspondence, morphological units, analogy, and whole-word recognition. Ziegler and Goswami (2005) went as far as to suggest that “it might even be the case that some of the most sophisticated processing architecture (e.g., two separate routes to pronunciation in the skilled reading system) may in fact only develop in English” (p. 20).

By making a distinction between the strategies employed by readers learning different orthographies to explain the cross-linguistic differences in reading accuracy and fluency, PGST has implications for the role of phonological and orthographic processing skills on reading development. If reading in orthographically inconsistent languages relies upon effective use of multiple recod-

---

George K. Georgiou and Rauno Parrila, Department of Educational Psychology, University of Alberta, Edmonton, Alberta, Canada; Timothy C. Papadopoulos, Department of Psychology, University of Cyprus, Nicosia, Cyprus.

Correspondence concerning this article should be addressed to George K. Georgiou, Department of Educational Psychology, 6-102 Education North, University of Alberta, Edmonton, Alberta T6G 2G5, Canada. E-mail: georgiou@ualberta.ca

ing strategies, skill at phonological awareness and orthographic processing should be important for reading development because decoding in orthographically inconsistent languages depends not only on the recognition of single graphemes and the retrieval of their corresponding sounds but also on the recognition of bigger grain size units, such as rimes and their corresponding sound units. Whereas phonological awareness is *sine qua non* for the former, orthographic knowledge is also needed for the latter (Share, 1995). Similarly, if reading in consistent orthographies relies on a grapheme–phoneme recoding strategy, then phonological short-term memory should play a larger role in reading in these languages because the phonological information of each grapheme must be available for blending and word naming to be successful.

### RAN and Reading Fluency

The PGST does not support predictions regarding the role of RAN on reading. However, some of the current theories of RAN seem to lead to different predictions regarding how orthographic consistency should affect RAN's contribution to reading development. If RAN measures the ability to access and retrieve phonological representations from long-term memory, as suggested by Torgesen, Wagner, and their colleagues (e.g., Torgesen, Wagner, Rashotte, Burgess, & Hecht, 1997; Wagner & Torgesen, 1987; Wagner et al., 1997), then it should be more strongly related to reading in consistent orthographies because the phonological representation of each grapheme should be retrieved quickly enough for the grapheme–phoneme recoding strategy to be effective. On the contrary, if RAN measures the ability to form orthographic representations, as suggested by Bowers and her colleagues (Bowers, Golden, Kennedy, & Young, 1994; Bowers, Sunseth, & Golden, 1999; Sunseth & Bowers, 2002), then RAN should be more important for reading in orthographically inconsistent languages because for the larger grain size unit strategies to succeed the orthographic information of those units must be developed.

Similarly, PGST does not explain the mechanism by which reading fluency is accomplished. Wimmer (2006) criticized PGST for providing a rather time-limited perspective because in consistent orthographies children achieve high levels of accuracy after a few months of reading instruction, and the main characteristic of further development is not reading accuracy but reading fluency. If children in consistent orthographies do not switch reading strategies depending on the type of reading task (accuracy or fluency), then fluency can be achieved only if grapheme–phoneme decoding strategy use becomes faster. If this inference is correct, then orthographic processing skills should not predict fluency development any more than accuracy development, whereas RAN should be a prominent predictor of reading fluency, an argument consistent with the data (e.g., Caravolas, 2006; de Jong & van der Leij, 1999). In contrast, in inconsistent orthographies fluency can be achieved only if multiple unit size strategies become more reliable and readily available; this would likely demand the joint contribution of both RAN and orthographic processing skills.

Because of the recency of the PGST and the RAN theories, the above predictions have not been directly tested. In the rest of this introduction, we review indirect evidence for the predictions, derived mainly from studies using either a sample learning to read a single language or from studies using a limited set of predictors in a multilanguage sample.

### Orthographic Consistency and Phonological Processing

Cross-sectional and longitudinal studies conducted in languages varying in orthographic consistency have produced conflicting findings as to the importance of each one of the phonological processing skills in reading acquisition. In English-speaking children, the contribution of phonological awareness appears to remain strong through elementary school, whereas the contribution of RAN appears to be time-limited and dependent on the type of RAN tasks used (Letter and Digit Naming vs. Color and Object Naming) and the reading competence of the children (e.g., Compton, 2003; Georgiou, Parrila, & Kirby, 2006; Meyer, Wood, Hart, & Felton, 1998; Schatschneider, Fletcher, Francis, Carlson, & Foorman, 2004). Finally, conflicting findings have been reported regarding the contribution of phonological short-term memory. For example, Swanson and his colleagues (e.g., Swanson & Alexander, 1997; Swanson & Howell, 2001) reported significant contributions of phonological short-term memory on reading, whereas others have shown that it is only weakly related to reading when considered along with phonological awareness and RAN (e.g., Parrila et al., 2004; Torgesen et al., 1997).

The majority of studies conducted in orthographically consistent languages have shown that phonological awareness either may not be an important predictor of reading (e.g., Aarnoutse, van Leeuwe, & Verhoeven, 2005; Harris & Giannouli, 1999; Holopainen et al., 2001) or may be important but only during the first 1 or 2 years of schooling (e.g., de Jong & van der Leij, 2002; Landerl & Wimmer, 2000; Leppänen, Niemi, Aunola, & Nurmi, 2006; Papadopoulos, 2001). Researchers hypothesized that the effect of consistent spelling–sound correspondences is sufficiently powerful to secure children's phonological recoding skills after a few months of reading experience, regardless of their prereading levels of phonological awareness (e.g., Caravolas, 2006; Porpodas, 1999; Wimmer, Landerl, & Schneider, 1994). Accordingly, several studies have suggested that RAN plays a more prominent role than phonological awareness in predicting reading development in consistent orthographies (e.g., de Jong & van der Leij, 1999, 2002; Mayringer, Wimmer, & Landerl, 1998; van den Bos, 1998; Wimmer, Mayringer, & Landerl, 2000). Finally, phonological short-term memory appears to play a trivial role in learning to read in orthographically consistent languages (e.g., de Jong & van der Leij, 1999; Nikolopoulos, Goulondris, Hulme, & Snowling, 2006).

The few studies that directly compared children learning different alphabetic languages have also provided mixed findings. Patel, Snowling, and de Jong (2004) compared the predictors of reading ability in English and Dutch and found that “the concurrent predictors of reading in English and Dutch were *strikingly similar* [italics added]” (p. 793). Phoneme deletion was a significant predictor of individual differences in reading, but RAN was not. The authors concluded that phonological awareness is a predictor of individual differences in reading skill in both transparent and opaque orthographies. However, the language by phoneme deletion interaction accounted for significant variance in word-reading accuracy after the effects of language, age, vocabulary, phoneme deletion, and RAN were controlled; follow-up analyses revealed that phoneme deletion was a significant predictor of word-reading accuracy only for the English children.

Mann and Wimmer (2002) examined the predictors of reading in English and German. At the end of kindergarten, Grade 1, and



Grade 2, children were given two tests of phonological awareness (Phoneme Identity Judgment and Phoneme Elision), RAN-Colors, letter identification, and short tests of word and nonword-reading accuracy and speed. Regression analyses showed that the only significant predictor of both reading accuracy and speed in English was phonological awareness. In German, there were no significant predictors of reading accuracy, whereas RAN was the only significant predictor of reading speed. Although Mann and Wimmer's (2002) results may reflect genuine differences in the extent to which reading acquisition relies on different processes, there are notable methodological problems that may compromise this conclusion. First, lack of variability in word and nonword-reading accuracy scores was evident in the German sample. Second, there was lack of variability in Phoneme Identity Judgment, a task used to create a composite phonological awareness score, which was, in turn, used to predict reading accuracy and speed.

Recently, Caravolas, Vólin, and Hulme (2005) suggested that when phonological awareness is measured with sufficiently difficult tasks, a significant contribution of phonological awareness on reading ability in regular orthographies can be detected even with older children. Caravolas et al. (2005) examined the effect of phonological awareness on reading and spelling in a regular orthography (Czech) and in an opaque orthography (English) in a group of normally developing children in Grades 2 to 5 (Czech sample) and Grades 2 to 7 (English sample). Phonological awareness, measured by Phoneme Elision and spoonerisms tasks, was a significant predictor of reading ability in both Czech and English.

To summarize, studies have provided contradictory findings on the contribution of phonological processing skills on reading across alphabetic languages that vary in orthographic consistency. The selection of tasks may moderate the relationship between phonological awareness, RAN, and reading. First, many phonological awareness tasks, such as rhyme awareness or phoneme judgment, are likely too easy for Grade 1 students in orthographically consistent languages, producing a ceiling effect. In contrast, a task such as Phoneme Elision is more difficult and is more likely to create variability in older readers. Second, alphanumeric RAN tasks (Digit and Letter Naming) in many single-language studies are stronger predictors of reading than nonalphanumeric RAN tasks (Color and Object Naming; e.g., Bowey, McGuigan, & Ruschena, 2005; Compton, 2003; Felton & Brown, 1990). Yet, no cross-linguistic studies used alphanumeric RAN tasks, possibly underestimating the RAN-reading relationship.

It is also possible that the differences found in the predictive value of the phonological processing skills may be partly due to the measures used to assess reading ability. Research in English-speaking populations has primarily focused on the prediction of reading accuracy, whereas research conducted in orthographically consistent languages has primarily focused on the prediction of individual differences in reading speed. Wolf and Bowers (1999) argued that phonological awareness is more strongly related to word-reading accuracy, and RAN is more strongly related to reading fluency (see also Bowers, 1995; Katzir et al., 2006; Manis, Doi, & Bhadha, 2000; Savage & Frederickson, 2005). Thus, to the extent that phonological awareness and RAN are differentially related to specific types of reading outcomes, then the use of reading-speed measures in consistent orthographies might have accentuated the role of RAN and the use of reading-accuracy

measures in inconsistent orthographies might have accentuated the role of phonological awareness.

### Orthographic Consistency and Orthographic Processing

The role of orthographic processing in predicting reading development is not yet fully understood. Confusion arises because of differences in the way orthographic processing has been conceptualized and operationalized (Burt, 2006; Hagiliassis, Pratt, & Johnston, 2006). Impressively, Wagner and Barker (1994) provided 11 definitions of orthographic processing used in research. For example, Stanovich and West (1989) defined orthographic processing as "the ability to form, store, and access orthographic representation(s)" (p. 404); and Perfetti (1984) defined it as "the knowledge a reader has about permissible letter patterns" (p. 47). In this study, we define orthographic processing as children's sensitivity to the orthographic structure of words.

Several studies have demonstrated orthographic processing's impact on reading acquisition in English (e.g., Badian, 2001; Cunningham, Perry, & Stanovich, 2001; Cutting & Denckla, 2001; Holland, McIntosh, & Huffman, 2004; Torgesen et al., 1997). For example, Torgesen et al. (1997) showed that orthographic processing accounted for a significant amount of unique variance in Grades 4 and 5 word-reading accuracy and reading comprehension.

Related studies conducted in other languages are rare, are confounded by bilingualism, and have produced conflicting results. For example, Arab-Moghaddam and Sénéchal (2001) examined the effects of phonological and orthographic processing skills on reading in English and Persian, an orthographically consistent language. Persian-speaking children in Grades 2 and 3 who had lived in English-speaking Canada for an average of 4 years were tested on word reading in English and Persian. Investigators found that the predictors of reading performance were similar across languages: Phonological and orthographic processing skills each accounted for unique variance in word reading in English and Persian once other related variables were controlled. Importantly, in both languages, the amount of unique variance accounted for by orthographic processing was twice the amount accounted for by phonological processing.

Geva, Wade-Woolley, and Shany (1993) studied children learning to read both English and vowelized Hebrew, which has almost perfect grapheme-to-phoneme correspondences. Although both phonological and orthographic processing predicted reading acquisition in English, only phonological skills predicted reading acquisition in Hebrew. These contrasting findings suggest that whereas orthographic processing contributes significantly to reading acquisition in English, its role is unclear in consistent orthographies.

### Overview of the Current Study

Several researchers have argued that in order for them to have a comprehensive understanding of the mechanisms involved in reading (e.g., Aro & Wimmer, 2003; Goswami, 1999; Harris & Hattano, 1999), models of reading development should be tested across languages varying in orthographic consistency. In response, the current study sought to compare concurrently and longitudinally the relative importance of phonological processing (phonological awareness, phonological memory, and RAN) and ortho-



graphic processing in predicting word-reading accuracy and fluency in children learning to read an orthographically consistent language (Greek) and in children learning to read an orthographically inconsistent language (English).

Within the family of alphabetic scripts, Greek provides an interesting contrast to English because of its high degree of regularity for reading. Although in Greek there are no statistical estimations of the degree of regularity, the almost perfect one-to-one correspondence between its graphemes and phonemes (see Porpodas, 2004, for a description of a few exceptions) and the predominance of open consonant-vowel syllables render Greek orthography consistent.

## Method

### Participants

Letters describing the study were sent to parents of 161 Grade 1 English-speaking Canadian and 92 Grade 1 Greek-speaking Cypriot children. One hundred thirty-two English-speaking children (70 girls and 62 boys, mean age = 79.48 months,  $SD = 3.98$ ) and 75 Greek-speaking children (42 girls and 33 boys, mean age = 82.20 months,  $SD = 3.33$ ), whose parents consented to participate in the study, were followed from Grade 1 to Grade 2. They were all native speakers of English and Greek, respectively. The Canadian children were assessed in April/May of Grade 1 and January/February of Grade 2, whereas the Cypriot children were assessed in April/May of Grade 1 and Grade 2, respectively. By Grade 2 the sample consisted of 110 English-speaking children (59 girls and 51 boys, mean age = 79.52 months,  $SD = 4.01$ ) and 70 Greek-speaking children (40 girls and 30 boys, mean age = 83.06 months,  $SD = 3.35$ ). Twenty-two English-speaking children (16.6% of the sample) and 5 Greek-speaking children (6.6% of the sample) withdrew from the study. To examine if the performance of the children who withdrew from the study differed significantly from the rest of the children, we performed  $t$  tests on their Grade 1 performance scores. None of the  $t$  tests reached significance for either English-speaking (all  $ps > .12$ ) or Greek-speaking (all  $ps < .08$ ) children. All the subsequent analyses were conducted with the children who were assessed at both measurement points. The children in both countries were coming mostly from families of middle-to-upper-middle socioeconomic status. None of the children participating in this study were identified as having learning, emotional, or sensory disabilities.

### Reading Instruction

Formal instruction in reading in Cyprus, where the data for the Greek-speaking children were collected, begins at the start of Grade 1, when children are, on average, 6 years of age. It emphasizes grapheme-phoneme correspondences or other letter combination patterns (i.e., digraphs). The whole language approach is also in use with the intention of building sight vocabulary. Grammatical and syntactic rules are introduced at the end of Grade 1 and are taught systematically from Grade 2 onward. The method of reading instruction in Alberta, where the data for the English-speaking children were collected, also emphasizes both grapheme-phoneme correspondences and whole-word recognition strategies. This combined method of reading instruction is known as the

Balanced Literacy Program (see Sénéchal & LeFevre, 2002; Sénéchal, LeFevre, Thomas, & Daley, 1998).

## Materials

Due to the cross-linguistic nature of the study, we decided to use those phonological awareness, phonological memory, and RAN measures that have been shown to be robust predictors of early reading in both English and Greek and that have similar or comparable versions for both languages. Similarly, we selected the orthographic, nonword-reading, and reading-fluency tasks in English for which comparable tasks could be developed in Greek.

### Phonological Awareness

In English, Elision, adopted from the Comprehensive Test of Phonological Processes (CTOPP; Wagner, Torgesen, & Rashotte, 1999), was used to measure phonological awareness. There were 3 practice items and 24 test items. Four test items required the participant to say the word without saying one of the syllables, and the remaining 20 items required the participant to say a word without saying a designated sound in the word. The position of the phoneme to be removed varied across those 20 items; in eight cases it involved the initial phoneme (e.g., *farm* without the /f/ is *arm*); in six cases, the medial phoneme (e.g., *winter* without the /t/ is *winner*), and in six cases, the final phoneme (e.g., *sheep* without the /p/ is *she*). Testing was discontinued after three consecutive errors. A participant's score was the number of correct items. The split-half reliability coefficient in our English-speaking sample was .92.

The Greek version of Elision had the same number of items as the English task. However, given that the items after the deletion of a phoneme must produce a real word and that, in contrast to the majority of the English words, the typical Greek words are two or more syllables, complete matching of the words in the two languages was impossible. Instead, we devised a list of items that required the children to delete the same sound (initial, medial, final) from the word as it was required from the English-speaking children. Specifically, four test items required the participant to say the word without saying one of the syllables (e.g., λεμόνι /lemoni/ [lemon] without the /le/ is μόνη /moni/ [alone]), eight test items required the participant to delete the initial phoneme (e.g., πόλη /poli/ [town] without the /p/ is όλοι /oli/ [all]); six test items required the participant to delete the medial phoneme (e.g., δίνω /dino/ [give] without the /n/ is δύο /dio/ [two]), and six test items required the participant to delete the final phoneme (e.g., ζώα /zoa/ [animals] without the /a/ is ζω /zo/ [live]). Testing was discontinued after three consecutive errors. A participant's score was the number of correct items. The split-half reliability coefficient of Phoneme Elision in our Greek-speaking sample was .91.

### Naming Speed

**Color Naming.** This task required participants to state as quickly as possible the names of five colors (blue, black, green, red, or yellow). The colors were presented on a laptop computer screen and arranged randomly in five rows with ten colors per row on two separate pages. Prior to beginning the timed naming, each participant was asked to name the colors in a practice trial to



ensure familiarity. The two pages were timed separately. Wolf and Denckla (2005) reported test-retest reliability of Color Naming to be .90. Color Naming in Greek was administered the same way as in English. The corresponding names of colors in Greek are μπλε (*mble*) for blue, μαύρο (*mavro*) for black, πράσινο (*prasino*) for green, κόκκινο (*kokkino*) for red, and κίτρινο (*kitrino*) for yellow. The mean phoneme length for the Greek color names was 5.6, whereas for the English it was 3.6. This difference was significant,  $t(8) = 2.54, p < .05$ .

**Digit Naming.** This task was adopted from CTOPP (Wagner et al., 1999). This RAN task consisted of a set of six digits (4, 7, 8, 5, 2, 3) that were displayed in random sequence six times for a total of 36 stimuli. Subjects were asked to name the digits from left to right as quickly as possible. The total time to completion was recorded. Before naming the 36 digits, each participant was asked to name the six digits in a practice trial to ensure familiarity. Wagner et al. (1999) reported test-retest reliability of .91 for Digit Naming for children ages 5 to 7 years. Digit Naming in Greek was administered in the same way as in English. The corresponding names of digits in Greek are τέσσερα (*tessera*) for four, επτά (*epta*) for seven, οκτώ (*okto*) for eight, πέντε (*pende*) for five, δύο (*dio*) for two, and τρία (*tria*) for three. The mean phoneme length for the Greek digit names was 4.8, whereas for the English it was 3.6. This difference was not significant,  $t(10) = 1.43, ns$ .

### Phonological Memory

Forward Digit Span from the Wechsler Intelligence Scale for Children-Third Edition (WISC-III; Wechsler, 1992) was used to assess phonological short-term memory in both language groups. The strings of digits were presented orally with a time interval of about 0.5 s between each digit. The child had to repeat the digits in each string in correct order. The strings started with only two digits, and one digit was added for each new digit string. The task was terminated when the child failed both trials of a given length. The performance score was the number of digit strings that the child could accurately provide. Split-half reliability coefficient for this task in our English-speaking sample was .69, whereas the corresponding reliability coefficient in our Greek-speaking sample was only .51.

### Orthographic Processing

The Orthographic Choice task was adapted from the work of Olson and colleagues (e.g., Olson, Forsberg, Wise, & Rack, 1994; Olson, Wise, Conners, Rack, & Fulker, 1989). The students viewed pairs of letter strings that sounded alike (e.g., rain - rane) and were asked to circle the one that was spelled correctly. Thirty pairs of phonologically similar letter strings were presented to the children on a sheet of paper. An individual's score was the number of correctly circled real words. Cronbach's alpha for Orthographic Choice in our English-speaking sample was .80. The Orthographic Choice test in Greek had the same number of items as in English. However, given that irregularity in Greek orthography can be found only in the spelling of the phonemes /o/, /i/, and /ε/, the selected real words and their pseudohomophones differed only in the way that the aforementioned phonemes were represented (e.g., /σχολείο/ [school] versus /σχολβο/). Cronbach's alpha in our Greek-speaking sample was .87.

### Nonword Reading

Form H Word Attack from Woodcock Reading Mastery Tests-Revised (WRMT-R; Woodcock, 1998) was used to assess decoding in Grades 1 and 2. Participants were asked to read 45 nonwords presented on a laptop screen as if they were real English words. The number of characters in the English Word Attack task was 215. Testing was discontinued after six consecutive errors. A participant's score was the number of items correct. Split-half reliability coefficient in our sample was .94 for Grade 1 and .93 for Grade 2, respectively.

Similar to its original version in English, the Greek Word Attack task consisted of 45 pronounceable nonwords that were derived from real words after changing two or three letters (either by substituting them or using them backwards). It was originally piloted in a study by Papadopoulos and Georgiou (2000) and was later used in other studies, showing satisfactory psychometric properties in its Greek version with both typically developing populations (e.g., Papadopoulos, 2001) and children exhibiting reading difficulties (e.g., Papadopoulos, Chalarambous, Kanari, & Loizou, 2004). The number of characters in the Greek Word Attack task was 218. Testing was discontinued after six consecutive errors. A participant's score was the number of items correct. Split-half reliability coefficient in our sample was .94 in Grade 1 and .87 in Grade 2, respectively.

### Reading Fluency

Gray Oral Reading Test (GORT; Wiederholt & Bryant, 2001) was used to assess subjects' text-reading fluency. The participants were asked to read as fast and as accurately as possible two short stories whose content was familiar to both North American and Cypriot children. The reading comprehension questions that follow each story were not administered because it was not the intention of this study to examine reading comprehension. Children's score was the total time taken to read the two short stories. Wiederholt and Bryant (2001) reported test-retest reliability for GORT to be .93. The Greek version of this task was a translation/back translation of the English GORT. Thus, although attention was paid to ensure that the words in the translated texts were of the same difficulty as in English, passage length across languages was dissimilar. The English GORT contained 61 words for a total of 232 characters, whereas the Greek GORT contained 53 words for a total of 281 characters.

Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999) was used to assess subjects' word-reading fluency. The child was given a list of 104 words, divided into four columns of 26 words each, and asked to read them as fast as possible. A short, 8-word practice list was presented first. The number of words read correctly and the number of errors made within a 45-s time limit was recorded. The score was the number of words read correctly. Torgesen et al. (1999) reported test-retest reliability of .95 for ages 6 to 9 years. The Greek version of this task had the same format as the English one. It consisted of 104 words beginning with one-syllable words and ending with three-syllable words. However, the Greek version of the task had longer words compared with the English task. More specifically, although the Greek TOWRE contained 644 characters, the English TOWRE contained 607 characters.

### Procedure

In Grade 1, the children were administered measures of phonological awareness, rapid-naming speed, phonological memory, orthographic processing, word decoding, and reading fluency. In Grade 2, only the dependent measures of word decoding and reading fluency were administered. All participants were tested individually in their respective schools during school hours by trained experimenters. Testing in Grade 1 was divided into two sessions lasting roughly 30 to 40 min. Session A consisted of Color Naming, Elision, Digit Naming, and Word Attack. Session B consisted of Digit Span, Orthographic Choice, TOWRE, and GORT. Half of the participants in each language group received first Session A, whereas the other half received first Session B. The order of the tasks within each session was fixed.

### Statistical Analysis

To examine the contribution of different predictors on reading accuracy and fluency both concurrently and longitudinally, we used structural equation modeling (SEM). Maximum likelihood estimation procedures were used to analyze the variance/covariance matrix of the observed variables using AMOS 7.0 (Arbuckle, 2006). To evaluate model fit, chi-square values and a set of fit indexes were used as follows: (a) the comparative fit index (CFI); (b) the goodness of fit index (GFI); and (c) the root-mean-square error of approximation (RMSEA). Nonsignificant chi-square, CFI, and GFI indices above .95 suggest model acceptance (Hu & Bentler, 1999). RMSEA values below or at .05 indicate a close fit but values as high as .07 are regarded as acceptable (Browne & Cudeck, 1993).

Separate models were constructed with Word Attack, TOWRE, and GORT as the outcome variables. The first step was to estimate the fit of a baseline model, depicted in Figure 1, with all possible correlations between the predictor variables (Elision, Digit Span, RAN-Digits, and Orthographic Choice) and all possible paths from the predictor variables to the outcome variable present in Grades 1 and 2 (Word Attack, TOWRE, and GORT) separately for English and Greek. Age was included as a control variable, but it was not allowed to correlate with other predictor variables in the model. Given that the models are close to saturated, we expected that they would provide an almost perfect fit to the data (e.g., Hayduk,

1986). To increase the degrees of freedom and examine whether the most parsimonious well-fitting models are similar in the two languages, nonsignificant correlations and regression paths were dropped one at a time until all remaining paths in the models were significant. Age was retained in all models as a control variable.

Next, we examined the cross-linguistic differences by performing multigroup analyses. To increase the degrees of freedom in the tested models, only those correlations between the predictor variables and paths from the predictor variables to the outcome measures that were significant in one or both of the languages were retained. We then tested first the fit of a multigroup model in which no cross-language constraints were imposed. This was followed by testing the invariance of the regression paths in the two language groups by imposing equality constraints on the direct effects of the predictor variables on Word Attack, TOWRE, and GORT. Specifically, regression paths were constrained in a step-wise fashion in which the direct effects of Elision on the reading outcomes were constrained first. In testing for the invariance of the regression paths, we compared the chi-square value of the constrained model with that of the initial multigroup model in which no cross-language constraints were imposed. If the difference in chi-square values, given the difference in the degrees of freedom between the two models ( $df$  constrained –  $df$  unconstrained), was significant, then this indicated that the specific predictor was contributing in a different way to the outcome variable in the two languages. Next, we examined, one at a time, the invariance of the regression paths of Digit Span, RAN-Digits, and Orthographic Choice on the reading outcomes following the same procedure. Testing for invariance was performed in the same way as described above for Elision.

Finally, we obtained estimates of total effects, or the sum of direct (unmediated) and indirect (mediated) effects, on Grade 2 reading outcomes. The SEM analysis model displayed in Figure 1 was fitted to the data three times for each language, first with Grade 2 word decoding accuracy (Word Attack) as the dependent variable and then with word-reading fluency (TOWRE) and text-reading fluency (GORT) scores as the dependent variables, respectively.

## Results

### Preliminary Data Analysis

Descriptive statistics for the entire sample of English and Greek children are shown in Table 1. One English- and 5 Greek-speaking children failed to name all the colors and were thus not administered the RAN-Colors task. Similarly, 2 English- and 2 Greek-speaking children could not complete GORT in Grade 1. In order to examine if there were significant differences between the two groups on age, we performed  $t$  tests. The results revealed that the Greek-speaking children were significantly older than the English-speaking children both in Grade 1,  $t(178) = 6.13$ ,  $p < .001$ , Cohen's  $d = 0.94$ , and in Grade 2,  $t(178) = 7.02$ ,  $p < .001$ , Cohen's  $d = 1.21$ .

A closer look at the distributional properties of the variables revealed some problems. The rapid naming speed tasks were positively skewed. Log transformations were applied to reach normality. The distribution of Orthographic Choice was negatively skewed. Reflection plus square-root transformation was

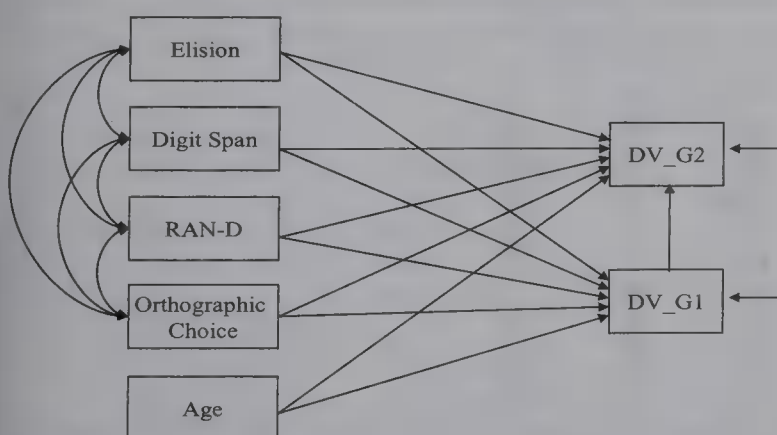


Figure 1. Model of relations between the predictor variables and the reading outcome in Grade 2. RAN-D = Rapid Automatized Naming-Digits; DV = dependent variable; G1 = Grade 1; G2 = Grade 2.



Table 1  
*Descriptive Statistics of All the Measures by Language*

Measure	English			Greek		
	<i>M</i>	<i>SD</i>	<i>N</i>	<i>M</i>	<i>SD</i>	<i>N</i>
Elision	11.79	5.77	110	14.51	5.45	70
RAN-Colors	1.29	0.44	109	1.33	0.34	65
RAN-Digits	0.61	0.22	110	0.55	0.14	70
Digit Span	6.30	1.55	110	6.04	1.18	70
OC	23.10	4.48	110	22.50	6.01	70
WAT_G1	16.13	9.33	110	38.66	6.50	70
WAT_G2	20.80	8.98	110	41.41	2.97	70
TOWRE_G1	35.48	15.37	110	31.19	11.39	70
TOWRE_G2	47.24	16.08	110	43.49	11.89	70
GORT_G1	77.06	65.90	108	82.64	38.16	68
GORT_G2	46.87	37.20	110	44.07	21.55	70

*Note.* RAN = Rapid Automatized Naming; OC = Orthographic Choice; WAT = Word Attack; TOWRE = Test of Word Reading Efficiency; GORT = Gray Oral Reading Test; G1 = Grade 1; G2 = Grade 2.

performed to reach normality. Despite the effectiveness of the transformations, some of the measures were still affected by the presence of outliers. More specifically, performances of 3 English-speaking and 4 Greek-speaking children were outliers on RAN-Digits and performances of 3 English-speaking and 2 Greek-speaking children were outliers on RAN-Colors. The outliers in both tasks were located at the high end of the distribution, indicating slow performance. In order to reduce the possible effect of outliers, we replaced their responses by a value equal to the next highest nonoutlier score plus one unit of measurement (Tabachnick & Fidell, 2001). The outlier's ranking within the distribution was preserved without deleting the datum or permitting its skewing influence. The transformed scores were used in all further analyses.

Table 2 displays the correlations between all the measures by language group. A similar pattern of correlations is observable in the two languages between the independent measures and reading. RAN-Digits was strongly related to reading fluency across languages and phonological awareness was strongly related to word

decoding in English. Orthographic Choice was strongly related to both word decoding and fluency measures across languages. Phonological memory was moderately related to different reading outcomes in English. The dependent measures of word decoding and fluency showed high stability from Grade 1 to Grade 2.

### *Predictors of Word Decoding*

To examine the predictors of word decoding, the baseline model (Figure 1) was fitted to the data with Grade 1 and Grade 2 Word Attack scores as the dependent variables. As was expected, the model fitted the data very well,  $\chi^2(4, N = 110) = 0.62, p = .892$ , CFI = 1.00, GFI = .99, RMSEA = .00, for English, and  $\chi^2(4, N = 70) = 1.44, p = .838$ , CFI = 1.00, GFI = .99, RMSEA = .00, for Greek, respectively. RAN-Digits was used as a measure of rapid naming speed because, in both languages, it produced generally larger correlations with the dependent variables than RAN-Colors. Next, we deleted one at a time the nonsignificant paths to produce the most parsimonious model in which Word Attack was the dependent variable in each language. Figure 2 shows the final models. The models provided a good fit to the data,  $\chi^2(7, N = 110) = 3.48, p = .837$ , CFI = 1.00, GFI = .99, RMSEA = .00, for English, and  $\chi^2(10, N = 70) = 5.45, p = .859$ , CFI = 1.00, GFI = .97, RMSEA = .00, for Greek, respectively, and accounted for a moderate proportion of the variance in Grade 1 (English:  $R^2 = .58$ ; Greek:  $R^2 = .41$ ) and a high proportion of the variance in Grade 2 for English ( $R^2 = .83$ ), but not for Greek ( $R^2 = .38$ ). There were two significant predictors of Grade 1 Word Attack in English: Elision and Orthographic Choice, and three significant predictors of Grade 1 Word Attack in Greek: Elision, Digit Span, and RAN-Digits. Not surprisingly, there was a strong autoregressive path from Word Attack in Grade 1 to Word Attack in Grade 2 in both languages. However, even after controlling for the autoregressive effect, Elision and RAN-Digits measured in Grade 1 were significant predictors of Word Attack in Grade 2 in English. In Greek, Word Attack in Grade 1 was the only significant predictor of Word Attack in Grade 2. It should be noted that given the strong autoregressive effect in both languages, the evaluation of

Table 2  
*Correlations Between All Measures by Language*

Measure	1	2	3	4	5	6	7	8	9	10	11	12
1. Age		.07	-.02	-.08	-.06	.08	.06	.03	.16	.17	-.18	-.18
2. Elision	.00		-.34**	-.35**	.27*	.31**	.53**	.35**	.36**	.32**	-.33**	-.36**
3. RAN-Colors	-.23*	-.39**		.51**	-.13	-.06	-.41**	-.34**	-.35**	-.34**	.40**	.35**
4. RAN-Digits	-.20	-.45**	.73**		-.14	-.44**	-.46**	-.41**	-.74**	-.64**	.71**	.59**
5. Digit Span	.00	.52**	-.17	-.24*		.39**	.41**	.25*	.26*	.30*	-.18	-.36**
6. OC	-.06	.57**	-.32**	-.43**	.44**		.38**	.32**	.66**	.69**	-.67**	-.67**
7. WAT_G1	-.04	.67**	-.20*	-.39**	.47**	.68**		.63**	.58**	.50**	-.53**	-.50**
8. WAT_G2	-.01	.70**	-.37**	-.49**	.44**	.67**	.89**		.48**	.51**	-.46**	-.47**
9. TOWRE_G1	.02	.64**	-.40**	-.56**	.49**	.82**	.77**	.80**		.87**	-.92**	-.83**
10. TOWRE_G2	.02	.57**	-.46**	-.64**	.40**	.72**	.70**	.79**	.86**		-.85**	-.89**
11. GORT_G1	.03	-.57**	.42**	.60**	-.44**	-.79**	-.70**	-.75**	-.91**	-.87**		.83**
12. GORT_G2	.01	-.59**	.42**	.66**	-.43**	-.69**	-.67**	-.75**	-.84**	-.91**	.89**	

*Note.* Correlations above the diagonal are from the Greek-speaking sample, whereas correlations below the diagonal are from the English-speaking sample. RAN = Rapid Automatized Naming; OC = Orthographic Choice; WAT = Word Attack; TOWRE = Test of Word Reading Efficiency; GORT = Gray Oral Reading Test; G1 = Grade 1; G2 = Grade 2.

\*  $p < .05$ . \*\*  $p < .01$ .

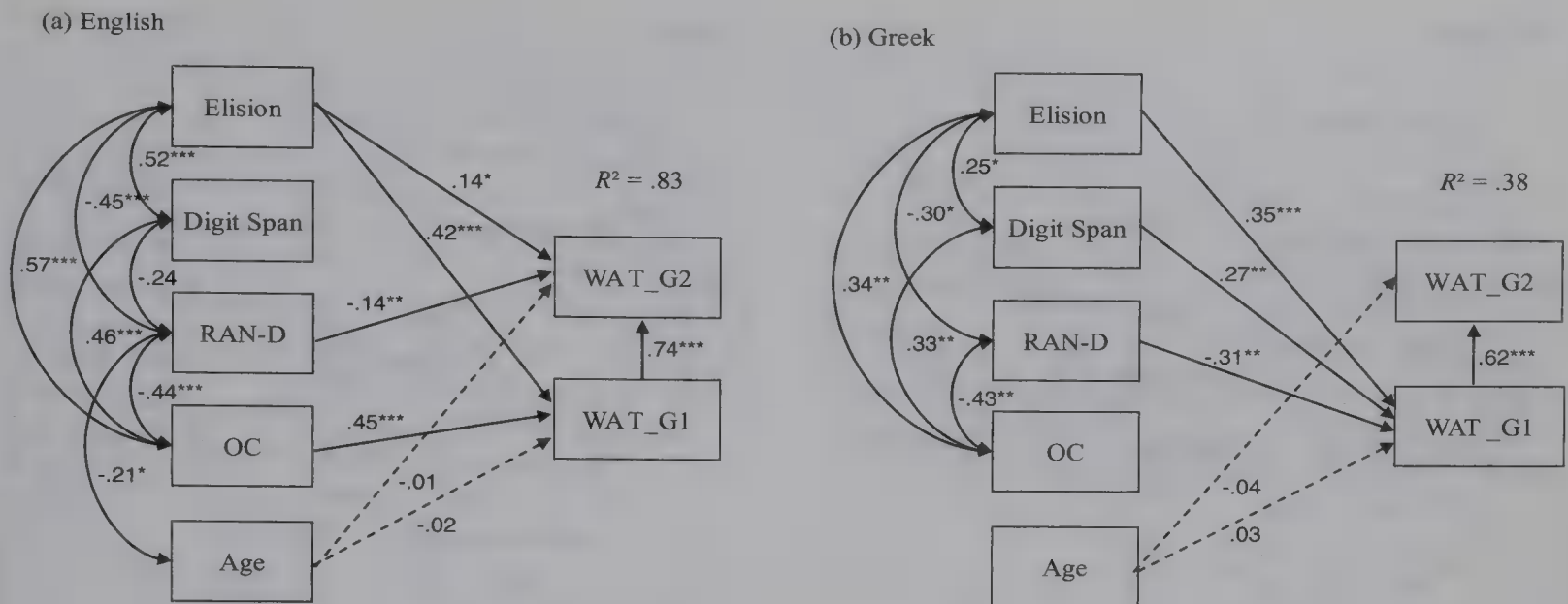


Figure 2. The baseline path model of predictors of Word Attack (WAT) in Grade 2 in English (a) and Greek (b). RAN-D = Rapid Automatized Naming-Digits; OC = Orthographic Choice; G1 = Grade 1; G2 = Grade 2. \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

other longitudinal predictors of Grade 2 Word Attack is highly conservative.

After establishing the most parsimonious model for each language separately, we used multigroup analyses to examine if there were any differences across the two languages in the significant predictors of Word Attack. The unconstrained multigroup model included all paths that were significant in either or both of the models shown in Figure 2 and fitted the data very well,  $\chi^2(10, N = 170) = 3.87, p = .953$ , CFI = 1.00, GFI = .99, RMSEA = .00.

Table 3 presents the changes in  $\chi^2$  values when predictors of interest were constrained to be equal across the languages. The fit of the multigroup model deteriorated significantly when the direct effects of Elision,  $\Delta\chi^2 = 8.79, p < .01$ , or Orthographic Choice,  $\Delta\chi^2 = 18.33, p < .001$ , on Grade 1 Word Attack were so constrained. In contrast, when the direct effects of Digit Span or RAN-Digits were constrained, there were no significant changes in model fit. For Word Attack in Grade 2, only constraining the direct effects of Elision,  $\Delta\chi^2 = 4.92, p < .05$ , and Word Attack in Grade

1,  $\Delta\chi^2 = 6.09, p < .05$ , resulted in a significantly poorer fitting model. Thus, these results indicate that Elision is a stronger predictor of Word Attack in English than in Greek both concurrently and longitudinally. Orthographic Choice, in turn, is a strong concurrent predictor of Grade 1 Word Attack in English only. Finally, the autoregressive effect of Grade 1 Word Attack on Grade 2 Word Attack was stronger in English than in Greek.

### Predictors of Word-Reading Fluency

The second set of analyses examined the predictors of TOWRE in Grade 1 and Grade 2. The baseline models fitted the data very well,  $\chi^2(4, N = 110) = 0.63, p = .890$ , CFI = 1.00, GFI = 1.00, RMSEA = .00, for English, and  $\chi^2(4, N = 70) = 1.48, p = .829$ , CFI = 1.00, GFI = .99, RMSEA = .00, for Greek, respectively. The most parsimonious model in each language is shown in Figure 3. The models provided a good fit to the data,  $\chi^2(7, N = 110) = 3.45, p = .840$ , CFI = 1.00, GFI = .99, RMSEA = .00, for

Table 3  
Changes in  $\chi^2$  After Constraining Paths to Be Equal Across Language Groups

Measure	Grade 1			Grade 2		
	WAT	TOWRE	GORT	WAT	TOWRE	GORT
EL <sub>E</sub> = EL <sub>G</sub>	8.79**	4.98*	—	4.92*	—	—
DS <sub>E</sub> = DS <sub>G</sub>	.12	—	—	—	—	—
RAN-D <sub>E</sub> = RAN-D <sub>G</sub>	1.33	8.15**	0.55	2.36	6.86**	4.29*
OC <sub>E</sub> = OC <sub>G</sub>	18.33**	10.93**	11.75**	—	0.29	3.38
WAT_G1 <sub>E</sub> = WAT_G1 <sub>G</sub>	—	—	—	6.09*	—	—
TOWRE_G1 <sub>E</sub> = TOWRE_G1 <sub>G</sub>	—	—	—	—	0.31	—
GORT_G1 <sub>E</sub> = GORT_G1 <sub>G</sub>	—	—	—	—	—	0.49
All direct effects equal	48.97***	23.36***	12.34*	24.72***	8.51*	10.21*
All indirect effects equal	—	—	—	31.15***	23.36**	12.59**

Note. WAT = Word Attack; TOWRE = Test of Word Reading Efficiency; GORT = Gray Oral Reading Test; EL = Elision; DS = Digit Span; RAN-D = Rapid Automatized Naming-Digits; OC = Orthographic Choice; E = English; G = Greek.

\* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .



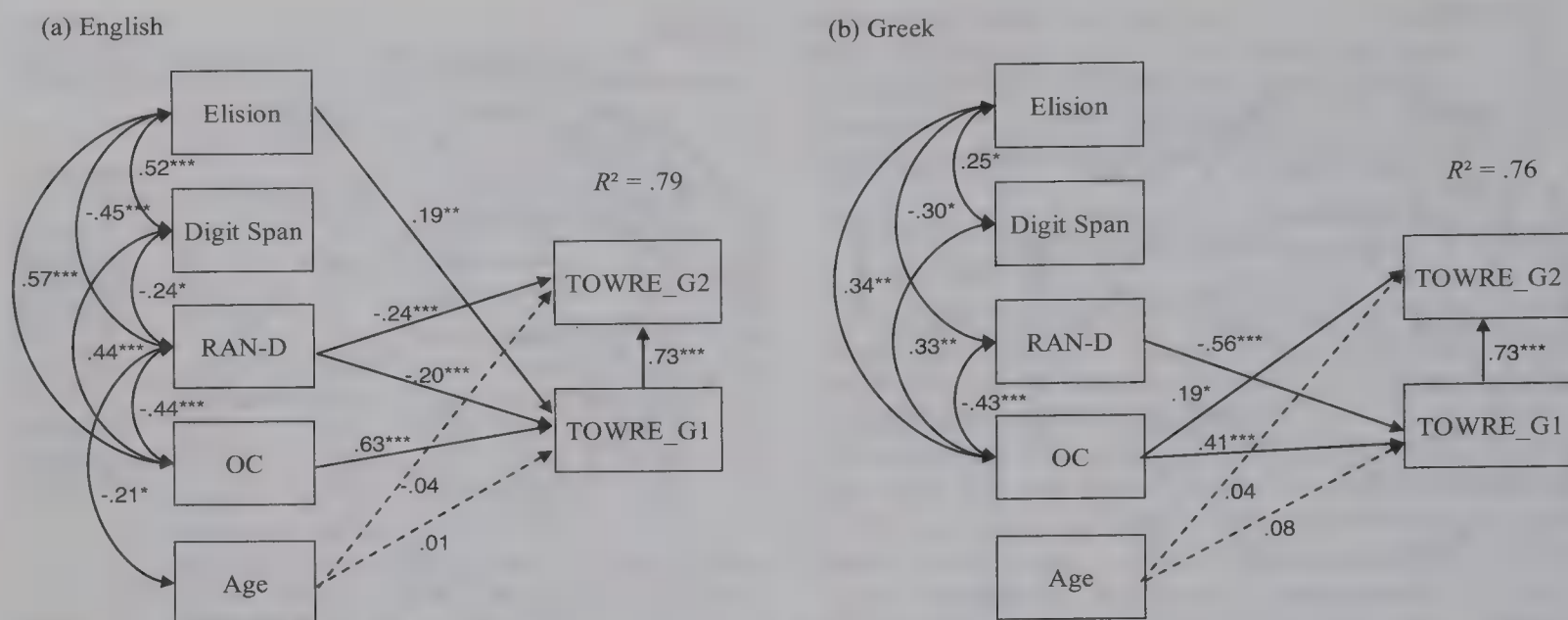


Figure 3. The baseline path model of predictors of the Test of Word Reading Efficiency (TOWRE) in Grade 2 in English (a) and Greek (b). RAN-D = Rapid Automatized Naming-Digits; OC = Orthographic Choice; G1 = Grade 1; G2 = Grade 2. \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

English, and  $\chi^2(10, N = 70) = 3.93, p = .950$ , CFI = 1.00, GFI = .98, RMSEA = .00, for Greek, respectively, and accounted for a high proportion of the variance in Grade 1 TOWRE (English:  $R^2 = .75$ ; Greek:  $R^2 = .68$ ) and Grade 2 TOWRE (English:  $R^2 = .79$ ; Greek:  $R^2 = .76$ ). There were three significant predictors of Grade 1 TOWRE in English: Elision, RAN-Digits, and Orthographic Choice; there were two unique predictors of Grade 1 TOWRE in Greek: RAN-Digits and Orthographic Choice. Longitudinally, there was a strong autoregressive path from Grade 1 TOWRE to Grade 2 TOWRE for both English and Greek. Despite the strong autoregressive effect, RAN-Digits in English and Orthographic Choice in Greek accounted for additional variance in Grade 2 TOWRE.

Next, we examined the cross-linguistic differences in the predictors of TOWRE. The unconstrained multigroup model,  $\chi^2(12, N = 170) = 5.23, p = .950$ , CFI = 1.00, GFI = .99, RMSEA = .00, included all paths that were significant in either or both of the models shown in Figure 3. When the direct effects of Elision,  $\Delta\chi^2 = 4.98, p < .05$ , RAN-Digits,  $\Delta\chi^2 = 8.15, p < .01$ , or Orthographic Choice,  $\Delta\chi^2 = 10.93, p < .01$ , on Grade 1 TOWRE were constrained to be equal across the two languages, the model fit deteriorated significantly. When the direct effect of RAN-Digits,  $\Delta\chi^2 = 6.86, p < .05$ , on Grade 2 TOWRE was constrained to be equal, the model fit also deteriorated significantly whereas the same was not true when the effect of Orthographic Choice or Grade 1 TOWRE was constrained. Thus, these analyses indicate that RAN-Digits exerts a significantly stronger effect in Greek than in English, and that Elision and Orthographic Choice were stronger predictors of Grade 1 TOWRE in English than in Greek.

#### Predictors of Text-Reading Fluency

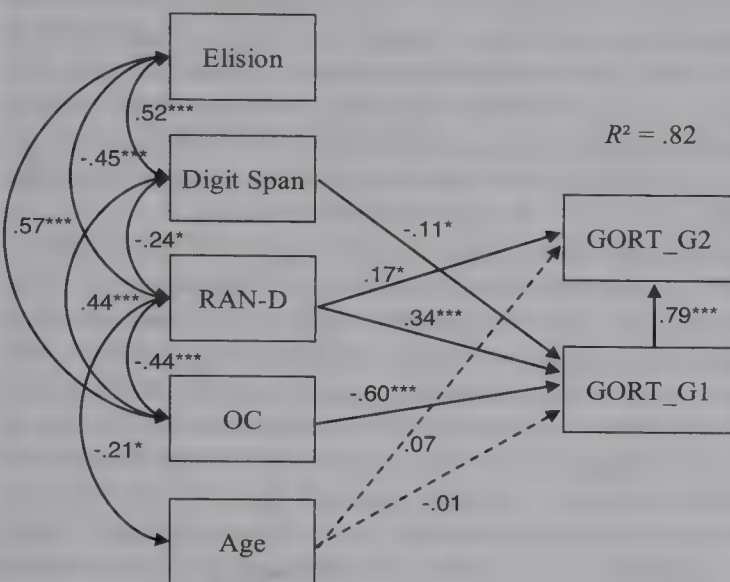
The third set of analyses examined the predictors of GORT in Grades 1 and 2. The baseline models fitted the data very well,  $\chi^2(4, N = 110) = 0.63, p = .990$ , CFI = 1.00, GFI = 1.00, RMSEA =

.00, for English, and  $\chi^2(4, N = 70) = 1.48, p = .829$ , CFI = 1.00, GFI = .99, RMSEA = .00, for Greek, respectively. Figure 4 presents the most parsimonious model for each language, each providing a good fit to the data,  $\chi^2(7, N = 110) = 4.45, p = .726$ , CFI = 1.00, GFI = .99, RMSEA = .00, for English, and  $\chi^2(10, N = 70) = 6.75, p = .748$ , CFI = 1.00, GFI = .98, RMSEA = .00, for Greek, respectively, and accounting for a high proportion of the variance in Grade 1 (English:  $R^2 = .75$ ; Greek:  $R^2 = .69$ ) and in Grade 2 (English:  $R^2 = .82$ ; Greek:  $R^2 = .71$ ). For both English and Greek, RAN-Digits and Orthographic Choice were the unique predictors of GORT in Grade 1. After controlling for the strong autoregressive path from GORT in Grade 1 to Grade 2, RAN-Digits in English and Orthographic Choice in Greek predicted unique variance in Grade 2 GORT.

Finally, we examined if there were any differences in the significant predictors of GORT across the two languages. The unconstrained multigroup model,  $\chi^2(12, N = 170) = 8.91, p = .711$ , CFI = 1.00, GFI = .99, RMSEA = .00, included all significant paths of the models shown in Figure 4. A significant change in model fit was observed when the effect of Orthographic Choice,  $\Delta\chi^2 = 11.75, p < .01$ , on Grade 1 GORT was constrained to be equal across the two languages. Similarly, constraining the path between RAN-Digits and Grade 2 GORT produced a significant change in the model fit,  $\Delta\chi^2 = 4.29, p < .05$ . Thus, similar to what we observed for TOWRE, Orthographic Choice was a better predictor of Grade 1 GORT in English than in Greek. In addition, RAN-Digits was a better predictor of Grade 2 GORT in English than in Greek.

To summarize, the SEM results demonstrated differences in the significant predictors of reading across languages. First, Elision appears to predict Word Attack better than it predicts TOWRE or GORT, and its contribution is greater in English than in Greek. Second, Digit Span was in most models a nonsignificant predictor of reading outcomes (with the exception of Grade 1 Word Attack in Greek and Grade 1 GORT in English). Third, RAN-Digits was

(a) English



(b) Greek

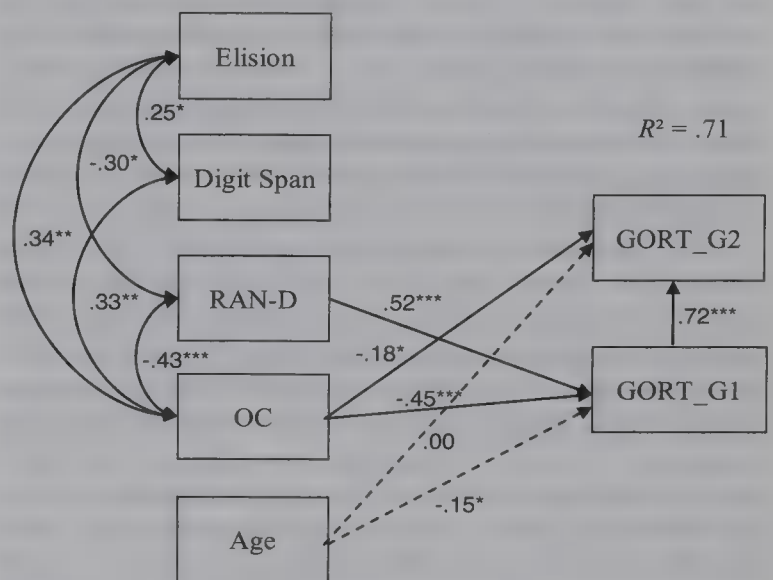


Figure 4. The baseline path model of predictors of the Gray Oral Reading Test (GORT) in Grade 2 in English (a) and Greek (b). RAN-D = Rapid Automatized Naming-Digits; OC = Orthographic Choice; G1 = Grade 1; G2 = Grade 2. \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

a stronger predictor of the reading fluency measures than of Word Attack. Notably, its effect on TOWRE in Grade 1 was stronger in Greek than in English. However, only in English did RAN-Digits have a direct effect on Grade 2 TOWRE and GORT. Fourth, Orthographic Choice was a stronger predictor of Grade 1 reading outcomes in English than in Greek. However, only in Greek did Orthographic Choice have a direct effect on Grade 2 TOWRE beyond the contribution of the autoregressor. Finally, many more similarities in the predictors were observed across languages in the reading fluency models than in the word decoding models.

#### Total Effects on Word Attack, TOWRE, and GORT in Grade 2

On several occasions, the above analyses indicated that the timing of the most significant effects may vary across the languages, creating an appearance of significant differences that may be short-lived. For example, in the TOWRE and GORT models for Greek-speaking children RAN-Digits predicted strongly Grade 1 reading performance but not Grade 2 performance, after the autoregressive effect was controlled. In contrast, similar models for the English-speaking children indicated that RAN-Digits predicted moderately both Grade 1 and Grade 2 reading outcomes. The opposite pattern, however, was observed for Orthographic Choice. These results leave open the possibility that in spite of initial significant differences between the languages, the total effect (the sum of direct and indirect effects) of RAN-Digits to Grade 2 reading fluency may be very similar.

Table 4 shows the standardized estimates of the total effects of Elision, RAN-Digits, Digit Span, and Orthographic Choice on Grade 2 Word Attack, TOWRE, and GORT in English and Greek. These estimates indicate that Elision had the highest impact on Word Attack in English. In contrast, RAN-Digits had the highest impact on Word Attack in Greek. In addition, Orthographic Choice contributed to Word Attack in English but not in Greek. The

reading fluency measures were clearly affected most by RAN-Digits and Orthographic Choice, whose contributions were comparable across languages. Digit Span made a small contribution to the dependent variables in both languages. Taken together with the SEM results, we can conclude that the differences in the predictors of reading in the two languages are more obvious for word decoding than for reading fluency measures. The differences are also more apparent in the contributions of Elision and Orthographic Choice than in the contributions of Digit Span and RAN.

#### Discussion

The primary objective of this study was to compare the concurrent and longitudinal predictors of word decoding and reading fluency in children learning to read the orthographically consistent language of Greek with children learning to read the orthographically inconsistent language of English. Consistent with the PGST, phonological awareness, measured with Elision, contributed sig-

Table 4

Total Effects of Elision, RAN-Digits, Digit Span, and Orthographic Choice on Grade 2 Word Attack, TOWRE, and GORT

Measure	English <sup>a</sup>			Greek <sup>b</sup>		
	WAT	TOWRE	GORT	WAT	TOWRE	GORT
Elision	.414	.085	-.103	.173	-.002	-.055
RAN-Digits	-.143	-.395	.426	-.298	-.404	.343
Digit Span	.024	.044	-.096	.129	.070	-.130
OC	.367	.484	-.405	.076	.467	-.439

Note. WAT = Word Attack, TOWRE = Test of Word Reading Efficiency; GORT = Gray Oral Reading Test; RAN = Rapid Automatized Naming; OC = Orthographic Choice.

<sup>a</sup>  $N = 110$ . <sup>b</sup>  $N = 70$ .



nificantly to the prediction of English Word Attack in Grades 1 and 2 and in Grade 1 TOWRE. Further, the multigroup analyses indicated that phonological awareness predicted Word Attack and TOWRE better in English than in Greek. These results suggest that reading development in consistent orthographies imposes fewer demands on phonological awareness than in inconsistent orthographies (e.g., Landerl & Wimmer, 2000; Mayringer et al., 1998; Mann & Wimmer, 2002; Wesseling & Reitsma, 2000).

RAN-Digits predicted significantly Grade 1 Word Attack only in Greek, but the difference with the English model was not significant. When reading fluency measures were the dependent variables, RAN-Digits significantly contributed in both languages. Notably, the effect of RAN-Digits on Grade 1 TOWRE was larger in Greek than in English, but only in English did RAN-Digits have a direct effect in Grade 2 fluency after controlling for the autoregressors. These results are in line with previous studies showing that RAN is a significant predictor of reading fluency (e.g., Bowers, 1995; Georgiou et al., 2006; Katzir et al., 2006; Savage & Frederickson, 2005; Schatschneider et al., 2004) and likely exerts its influence early on in reading development (e.g., Compton, 2003; de Jong & van der Leij, 2002; Wagner et al., 1997).

When we repeated the SEM analyses with RAN-Colors in the place of RAN-Digits, RAN-Colors predicted significantly word decoding and reading fluency in Greek, but not in English. Constraining the regression weights to be equal across languages resulted in a nonsignificant change in the  $\chi^2$  value. The fundamental question of why RAN-Digits is a stronger predictor of reading ability than RAN-Colors remains to be resolved. Recently, Bowey and colleagues (2005) showed that the robust association between alphanumeric RAN tasks (Digit and Letter Naming) and reading within their fourth-grade sample was largely mediated by phonological processing and that, relative to nonalphanumeric RAN tasks (Color and Object Naming), alphanumeric RAN better assesses an underlying phonological processing ability that is common to word-reading ability. Correlations in Table 2 indicate that for English-speaking children, RAN-Digits did indeed show stronger correlations with Elision than RAN-Colors. However, the same correlational pattern was not evident in the Greek-speaking sample. As such, the results of this study do not provide clear support for either of the two main theoretical accounts explaining why RAN is related to reading (Bowers & Wolf, 1993; Torgesen et al., 1997). If RAN was measuring the speed of access to phonological representations, it should be more strongly related to reading in Greek. However, RAN-Digits predicted Word Attack similarly across languages. Likewise, if RAN was measuring the ability to form high quality orthographic representations, it should be more important for reading in English. However, RAN-Digits was a stronger predictor of Grade 1 TOWRE in Greek than in English. What RAN tasks measure and why they are related to reading are clearly questions that warrant further investigation.

The third component of phonological processing, phonological memory, contributed significantly only when predicting Grade 1 word decoding in Greek. This result may reflect the fact that the last few items in Word Attack in Greek were relatively long and retaining phonological information in short-term memory would have been helpful for successful blending and naming to take place. Nevertheless, the results of multigroup analysis indicated that the differences between the regression weights in the two languages were not statistically significant. Thus, our results are in

line with those reported recently by Caravolas and her colleagues (2005). In their study, general IQ and Digit Span were the only nonsignificant predictors of reading speed and spelling in English and Czech. Likewise, there are a number of studies conducted in various languages showing that phonological memory is only weakly correlated with reading ability (e.g., de Jong & van der Leij, 1999; Dufva, Niemi, & Voeten, 2001; Muter & Snowling, 1998; Parrila et al., 2004; Scarborough, 1998).

Orthographic processing was a robust predictor of word decoding in English and of reading fluency in both languages. This finding is particularly important in light of arguments regarding the preferred strategy in word decoding across languages that vary in orthographic consistency (Ziegler & Goswami, 2005, 2006). Several researchers have argued that children who are learning to read an orthographically consistent language rely heavily on grapheme-phoneme decoding strategies (e.g., Aro & Wimmer, 2003; Ellis et al., 2004; Goswami, 2002; Havelka & Rastle, 2005; Wimmer & Goswami, 1994). In such languages, phonological recoding can reliably operate at the smallest grain size because the mapping of graphemes onto phonemes is unambiguous. In contrast, children who are learning to read an orthographically inconsistent language must use a variety of decoding strategies (e.g., Decker, Simpson, Yates, & Locker, 2003; Goswami, 2002; Goswami, Porpodas, & Wheelwright, 1997; Seymour, 2005). Our decoding accuracy findings support these arguments. In English, the best predictors of word decoding were phonological awareness and orthographic processing. In Greek, the best predictor of word decoding was RAN. Orthographic processing did not contribute significantly. The significant effect of orthographic processing on Word Attack in English may also reflect that many of the Word Attack items in English have common letter patterns that can be decoded as orthographic units rather than letter by letter.

To the extent that our orthographic processing measure captured the use of sublexical orthographic units, the results show that readers in English rely on both small and large grain size units to decode (e.g., Goswami, 1999; Goswami et al., 1997; Ziegler & Goswami, 2005). Importantly, although orthographic processing did not predict word decoding in Greek, it did predict significantly reading fluency in both Grades 1 and 2 even when the autoregressor's effect was controlled. This finding suggests that Greek-speaking children made use of their orthographic knowledge only when there was a need for a speeded response.

Our findings provide an extension to PGST (Ziegler & Goswami, 2005) by providing an account of how grain size theory may be used to explain the rapid attainment of reading fluency in consistent orthographies. Recently, de Jong (2006) and Wimmer (2006) criticized PGST for its explicit focus on the development of reading accuracy. Specifically, Wimmer (2006) argued that "learning to read (as viewed in PGST) is more or less equated with acquisition of recoding accuracy" (p. 447) and went on to propose that if PGST postulates a developmental trend from relying on small grain sizes to relying on larger ones, then it could accommodate reading fluency attainments.

However, it may also be the case that, with development, speed of processing small grain size units becomes faster and this, per se, leads to better performance in fluency tasks. If children learning to read a consistent orthography receive positive feedback on effective use of phonological recoding in word identification, then they may rely on the same grain size for efficient reading fluency. This,



in turn, would result in word length effects. Indeed, Ziegler, Perry, Jacobs, and Braun (2001) demonstrated that German adult readers exhibited strong length effects for words and nonwords compared with their English-speaking Australian counterparts. Prompted by this finding, Ziegler et al. concluded that "orthographic consistency determines not only the relative contribution of orthographic versus phonological codes within a given orthography, but also the preferred grain size of units that are likely to be functional during reading" (p. 379). However, evidence from eye-movement studies and word naming in Italian (a consistent orthography) challenges this conclusion (e.g., Barca, Burani, Di Filippo, & Zoccolotti, 2006; Orsolini, Fanari, Tosi, De Nigris, & Carrieri, 2006; Zoccolotti et al., 2005). For example, Spinelli et al. (2005) reported that Italian fourth-grade children did not show a length effect for reading words from three to five letters. In adult readers the effect remained absent for up to eight letters. The absence of a length effect in high frequency words suggests that during the development of reading a serial sublexical word-reading strategy is supplemented by a more parallel lexical strategy.

A third possibility is that children in consistent orthographies demonstrate flexibility in using different grain size units. In timed conditions when a response must be generated quickly, large grain size units are likely to be employed, whereas in untimed conditions when maximum accuracy is desirable, phonological recoding is likely to be employed. Our findings support the latter explanation. In word decoding, Greek-speaking children relied on small grain size units as indicated by the significant effect of phonological awareness. In contrast, in reading fluency tasks, Greek-speaking children relied on large grain size units as indicated by the significant effect of orthographic processing. Thus, Greek-speaking children may adjust the grain size units to match the task demands. When no speed is required there is a reliance on phonological recoding and when a speeded response is needed, bigger grain size units are employed. Even in consistent orthographies, two routes of pronunciation assembly may be necessary to describe the reading process, but the activation of each route relies upon the demands of the reading task.

### Limitations

Some limitations of this study are worth mentioning. First, our findings can be generalized only to the languages under investigation and for the ages we tested. Greek has been used by many researchers as an example of a transparent orthography (e.g., Goswami et al., 1997; Seymour, Aro, & Erskine, 2003), but there are still objections to this conceptualization (e.g., Miles, 2000). Because the purpose of this study was to compare reading processes in a consistent orthography with an inconsistent orthography, using Greek might not have been as optimal a comparison language as Finnish or Turkish, which are considered nearly 100% consistent.

Second, the reading accuracy and fluency tasks were not strictly matched in the two languages. Greek and English belong to different families of languages and have different orthographic and phonological characteristics. For example, there is a large body of short single-syllable words in English, whereas there is only a small number of such words in Greek. Given the number of single-syllable words used in existing reading tests in English (e.g., WRMT-R; Woodcock, 1998), it was not possible to construct

word-reading tasks in Greek that would be strictly parallel in terms of length and word frequency to the English ones. On the other hand, using more multisyllabic words in English likely will not create an equal task because of significant differences in the syllable structures between Greek and English.

Third, we should acknowledge the whole-word nature of the orthographic processing task used in our study. Grain sizes bigger than graphemes but smaller than words should also be tested. A task such as Word Likeness in which individuals are asked to select the letter string that looks like a real word in English from a pair of pronounceable nonwords (e.g., *filk* - *filv*) should be used in future studies. Such studies would provide more convincing evidence that orthographic knowledge (lexical and sublexical) predicts reading differently in languages varying in orthographic consistency.

Fourth, similar to many previous studies (e.g., Holland et al., 2004; Leppänen et al., 2006; Parrila et al., 2004; Schatschneider et al., 2004), we did not include any measures of intelligence. Although frequently used, controlling for intelligence can also be ill-advised unless the relationship between the chosen general ability measure and other predictors is well understood. As pointed out by Parrila et al. (2004), an additional problem arises when intelligence is operationalized as a vocabulary measure that is likely differentially related to various phonological processing abilities (e.g., Avons, Wragg, Cupples, & Lovegrove, 1998; Metsala, 1999). Furthermore, other cross-linguistic studies have indicated that IQ does not contribute significantly to reading speed (e.g., Caravolas et al., 2005). Finally, we made sure that the children who participated in this study were able to understand and execute the instructions and that none of them had diagnosed developmental disabilities.

The last limitation concerns the use of observed variables instead of latent variables in the SEM analyses. When relationships among latent variables are examined, the relationships are free of measurement error because the error can be estimated and removed. On the other hand, SEM analyses using observed variables assume that the measures have perfect reliability coefficients, which clearly is not the case. Particularly, Digit Span in Greek had low reliability, which may have attenuated the correlations with the reading outcomes. However, the results for Digit Span were so similar for the English- and Greek-speaking sample that low reliability is unlikely to be a major issue.

### Conclusions

Our findings add to a small but growing body of research that directly compares children learning to read in two different languages and suggest that both phonological and orthographic processing skills are important in early reading acquisition. However, at least at the early stages of reading development, the effects of them are moderated by the characteristics of the language, a conclusion that is based on the differential effects of phonological awareness and orthographic processing on word decoding and reading fluency measures. Although some of the differences might be short-lived (see Table 4), our findings challenge Patel and colleagues' (2004) suggestion that "to the extent that alphabetic languages map orthography to phonology, the predictors of reading performance are likely to be universal" (p.794). In contrast, we argue that the orthography that children are learning to read is an



important factor that needs to be taken into account when models of reading development are being generalized across languages. By considering the differences between language systems, cross-linguistic studies can be a powerful tool in our attempt to understand both universal and language-specific processes involved in reading acquisition.

## References

- Aarnoutse, C., van Leeuwe, J., & Verhoeven, L. (2005). Early literacy from a longitudinal perspective. *Educational Review and Research, 11*, 253–275.
- Arab-Moghaddam, N., & Sénéchal, M. (2001). Orthographic and phonological processing skills in reading and spelling in Persian/English bilinguals. *International Journal of Behavioral Development, 25*, 140–147.
- Arbuckle, J. L. (2006). *AMOS 7.0 user's guide*. Chicago: SPSS.
- Aro, M., & Wimmer, H. (2003). Learning to read: English in comparison to six more regular orthographies. *Applied Psycholinguistics, 24*, 621–635.
- Avons, S. E., Wragg, C. A., Cupples, L., & Lovegrove, W. J. (1998). Measures of phonological short-term memory and their relationship to vocabulary development. *Applied Psycholinguistics, 19*, 583–601.
- Badian, N. A. (2001). Phonological and orthographic processing: Their roles in reading prediction. *Annals of Dyslexia, 51*, 179–202.
- Barca, L., Burani, C., Di Filippo, G., & Zoccolotti, P. (2006). Italian developmental dyslexic and proficient readers: Where are the differences? *Brain and Language, 98*, 347–351.
- Barker, T. A., Torgesen, J. K., & Wagner, R. K. (1992). The role of orthographic processing skills on five different reading tasks. *Reading Research Quarterly, 27*, 335–345.
- Bowers, P. G. (1995). Tracing symbol naming speed's unique contributions to reading disability over time. *Reading and Writing: An Interdisciplinary Journal, 7*, 189–216.
- Bowers, P. G., Golden, J. O., Kennedy, A., & Young, A. (1994). Limits upon orthographic knowledge due to processes indexed by naming speed. In V. W. Berninger (Ed.), *The varieties of orthographic knowledge: Theoretical and developmental issues* (pp. 173–218). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Bowers, P. G., Sunseth, K., & Golden, J. (1999). The route between rapid naming and reading progress. *Scientific Studies of Reading, 3*, 31–53.
- Bowers, P. G., & Wolf, M. (1993). Theoretical links among naming speed, precise timing mechanisms and orthographic skill in dyslexia. *Reading and Writing: An Interdisciplinary Journal, 5*, 69–85.
- Bowey, J. A., McGuigan, M., & Ruschena, A. (2005). On the association between serial naming speed for letters and digits and word-reading skill: Towards a developmental account. *Journal of Research in Reading, 28*, 400–422.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways in assessing model fit. In K. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.
- Burt, J. S. (2006). What is orthographic processing skill and how does it relate to word identification in reading? *Journal of Research in Reading, 29*, 400–417.
- Caravolas, M. (2006, July). *The foundations of literacy in Czech: Not so different from English after all?* Paper presented at the 11th annual conference of the Society for the Scientific Studies of Reading, Vancouver, Canada.
- Caravolas, M., Vólin, J., & Hulme, C. (2005). Phoneme awareness is a key component of alphabetic literacy skills in consistent and inconsistent orthographies: Evidence from Czech and English children. *Journal of Experimental Child Psychology, 92*, 107–139.
- Compton, D. L. (2003). Modeling the relationship between growth in rapid naming speed and growth in decoding skill in first-grade children. *Journal of Educational Psychology, 95*, 225–239.
- Cunningham, A. E., Perry, K. E., & Stanovich, K. E. (2001). Converging evidence of the concept of orthographic processing. *Reading and Writing: An Interdisciplinary Journal, 14*, 549–568.
- Cutting, L. E., & Denckla, M. B. (2001). The relationship of rapid serial naming and word reading in normally developing readers: An exploratory model. *Reading and Writing: An Interdisciplinary Journal, 14*, 673–705.
- Decker, G., Simpson, G., Yates, M., & Locker, L. (2003). Flexible use of lexical and sublexical information in word recognition. *Journal of Research in Reading, 26*, 280–286.
- de Jong, P. F. (2006). Units and routes of reading in Dutch. *Developmental Science, 9*, 441–442.
- de Jong, P. F., & van der Leij, A. (1999). Specific contributions of phonological abilities to early reading acquisition: Results from a Dutch latent variable longitudinal study. *Journal of Educational Psychology, 91*, 450–476.
- de Jong, P. F., & van der Leij, A. (2002). Effects of phonological abilities and linguistic comprehension on the development of reading. *Scientific Studies of Reading, 6*, 51–77.
- Dufva, M., Niemi, P., & Voeten, M. J. (2001). The role of phonological memory, word recognition, and comprehension skills in reading development: From preschool to grade 2. *Reading and Writing: An Interdisciplinary Journal, 14*, 91–117.
- Ellis, N., Natsume, M., Stavropoulou, K., Hoxhallari, L., van Daal, V., Polyzoe, N., et al. (2004). The effects of orthographic depth on learning to read alphabetic, syllabic, and logographic scripts. *Reading Research Quarterly, 39*, 438–468.
- Felton, R. H., & Brown, I. S. (1990). Phonological processes as predictors of specific reading skills in children at risk for reading failure. *Reading and Writing: An Interdisciplinary Journal, 2*, 39–59.
- Frith, U. (1985). Beneath the surface of developmental dyslexia. In K. E. Patterson, M. Coltheart, & J. C. Marshall (Eds.), *Surface dyslexia* (pp. 301–330). London: Erlbaum.
- Georgiou, G., Parrila, R., & Kirby, J. R. (2006). Rapid automatized naming components and early reading acquisition. *Scientific Studies of Reading, 10*, 199–220.
- Geva, E., Wade-Woolley, L., & Shany, M. (1993). The concurrent development of spelling and decoding in different orthographies. *Journal of Reading Behaviour, 25*, 383–406.
- Goswami, U. (1999). The relationship between phonological awareness and orthographic representation in different orthographies. In M. Harris & G. Hatano (Eds.), *Learning to read and write: A cross-linguistic perspective* (pp. 134–156). Cambridge, England: Cambridge University Press.
- Goswami, U. (2002). Phonology, reading development, and dyslexia: A cross-linguistic perspective. *Annals of Dyslexia, 52*, 141–163.
- Goswami, U., Porpodas, C., & Wheelwright, S. (1997). Children's orthographic representations in English and Greek. *European Journal of Psychology of Education, 12*, 273–292.
- Hagiliassis, N., Pratt, C., & Johnston, M. (2006). Orthographic and phonological processes in reading. *Reading and Writing: An Interdisciplinary Journal, 19*, 235–263.
- Harris, M., & Giannouli, V. (1999). Learning to read and spell in Greek: The importance of letter knowledge and morphological awareness. In M. Harris & G. Hatano (Eds.), *Learning to read and write: A cross-linguistic perspective* (pp. 51–70). Cambridge, England: Cambridge University Press.
- Harris, M., & Hatano, G. (1999). Introduction: A cross-linguistic perspective on learning to read and write. In M. Harris & G. Hatano (Eds.), *Learning to read and write: A cross-linguistic perspective* (pp. 1–9). Cambridge, England: Cambridge University Press.
- Havelka, J., & Rastle, K. (2005). The assembly of phonology from print is



- serial and subject to strategic control: Evidence from Serbian. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 148–158.
- Hayduk, L. (1986). *Structural equation modeling with LISREL*. Baltimore, MD: Johns Hopkins University Press.
- Holland, J., McIntosh, D., & Huffman, L. (2004). The role of phonological awareness, rapid automatized naming, and orthographic processing in word reading. *Journal of Psychoeducational Assessment*, 22, 233–260.
- Holopainen, L., Ahonen, T., & Lyytinen, H. (2001). Predicting delay in reading achievement in a highly transparent language. *Journal of Learning Disabilities*, 34, 401–413.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1–55.
- Katzir, T., Kim, Y., Wolf, M., O'Brien, B. K., Lovett, M., & Morris, R. (2006). Reading fluency: The whole is more than the parts. *Annals of Dyslexia*, 56, 51–82.
- Landerl, K., & Wimmer, H. (2000). Deficits in phoneme segmentation are not the core problem of dyslexia: Evidence from German and English children. *Applied Psycholinguistics*, 21, 243–262.
- Leppänen, U., Niemi, P., Aunola, K., & Nurmi, J.-E. (2006). Development of reading and spelling Finnish from preschool to grade 1 and grade 2. *Scientific Studies of Reading*, 10, 3–30.
- Manis, F. R., Doi, L. M., & Bhadha, B. (2000). Naming speed, phonological awareness, and orthographic knowledge in second graders. *Journal of Learning Disabilities*, 33, 325–333.
- Mann, V., & Wimmer, H. (2002). Phoneme awareness and pathways into literacy: A comparison of German and American children. *Reading and Writing: An Interdisciplinary Journal*, 15, 653–682.
- Marsh, G., Friedman, M., Welch, V., & Desberg, P. (1981). A cognitive developmental theory of reading acquisition. In G. E. Mackinnon & T. G. Walker (Eds.), *Reading Research: Advances in Theory and Practice* (Vol. 3, pp. 199–221). New York: Academic Press.
- Mayringer, H., Wimmer, W., & Landerl, K. (1998). Phonological skills and literacy acquisition in German. In P. Reitsma & L. Verhoeven (Eds.), *Problems and interventions in literacy development* (pp. 147–161). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Metsala, J. (1999). Young children's phonological awareness and nonword repetition as a function of vocabulary development. *Journal of Educational Psychology*, 91, 3–19.
- Meyer, M. S., Wood, F. B., Hart, L. A., & Felton, R. H. (1998). Selective predictive value of rapid automatized naming in poor readers. *Journal of Learning Disabilities*, 31, 106–117.
- Miles, E. (2000). Dyslexia may show a different face in different languages. *Dyslexia*, 6, 193–201.
- Muter, V., Hulme, C., Snowling, M., & Stevenson, J. (2004). Phonemes, rimes vocabulary, and grammatical skills as foundations of early reading development: Evidence from a longitudinal study. *Developmental Psychology*, 40, 665–681.
- Muter, V., & Snowling, M. (1998). Concurrent and longitudinal predictors of reading: The role of metalinguistic and short-term memory skills. *Reading Research Quarterly*, 33, 320–337.
- Nikolopoulos, D., Goulondris, N., Hulme, C., & Snowling, M. J. (2006). The cognitive bases of learning to read and spell in Greek: Evidence from a longitudinal study. *Journal of Experimental Child Psychology*, 94, 1–17.
- Olson, R., Forsberg, H., Wise, B., & Rack, J. (1994). Measurement of word recognition, orthographic, and phonological skills. In G. R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities* (pp. 243–277). Baltimore, MD: Brookes.
- Olson, R., Wise, B., Conners, F., Rack, J., & Fulker, D. (1989). Specific deficits in component reading and language skills: Genetic and environmental influences. *Journal of Learning Disabilities*, 22, 339–348.
- Orsolini, M., Fanari, R., Tosi, V., De Nigris, B., & Carrieri, R. (2006). From phonological recoding to lexical reading: A longitudinal study on reading development in Italian. *Language and Cognitive Processes*, 21, 576–607.
- Papadopoulos, T. C. (2001). Phonological and cognitive correlates of word-reading acquisition under two different instructional approaches. *European Journal of Psychology of Education*, 16, 549–567.
- Papadopoulos, T. C., Charalambous, A., Kanari, A., & Loizou, M. (2004). Kindergarten intervention for dyslexia: The PREP remediation in Greek. *European Journal of Psychology of Education*, 19, 79–105.
- Papadopoulos, T. C., & Georgiou, G. K. (2000). Parameters of reading development in Greek language. In S. Georgiou, L. Kyriakides, & K. Christou (Eds.), *Contemporary research in educational studies* (pp. 241–248). Nicosia, Cyprus: University of Cyprus Press. [In Greek]
- Parrila, R. K., Kirby, J. R., & McQuarrie, L. (2004). Articulation rate, naming speed, verbal short-term memory, and phonological awareness: Longitudinal predictors of early reading development? *Scientific Studies in Reading*, 8, 3–26.
- Patel, T., Snowling, M. J., & de Jong, P. F. (2004). A cross-linguistic comparison of children learning to read in English and Dutch. *Journal of Educational Psychology*, 96, 785–797.
- Perfetti, C. A. (1984). *Reading ability*. New York: Oxford University Press.
- Porpodas, C. (1999). Patterns of phonological and memory processing in beginning readers and spellers of Greek. *Journal of Learning Disabilities*, 32, 404–416.
- Porpodas, C. (2004). Reading, spelling, and dyslexia in Greek: Research on the role of linguistic and cognitive skills. In I. Smythe, J. Everatt, & R. Salter (Eds.), *International book of dyslexia: A cross-language comparison and practice guide* (pp. 105–112). Chichester, England: Wiley.
- Savage, R., & Frederickson, N. (2005). Evidence of a highly specific relationship between rapid automatic naming of digits and text-reading speed. *Brain and Language*, 93, 152–159.
- Scarborough, H. S. (1998). Early identification of children at risk for reading disabilities: Phonological awareness and some other promising predictors. In B. Shapiro, P. Accardo, & A. Capute (Eds.), *Specific reading disability: A view of the spectrum* (pp. 75–119). Timonium, MD: York Press.
- Schatschneider, C., Fletcher, J. M., Francis, D. J., Carlson, C. D., & Foorman, B. R. (2004). Kindergarten prediction of reading skills: A longitudinal comparative analysis. *Journal of Educational Psychology*, 96, 265–282.
- Sénéchal, M., & LeFevre, J. (2002). Parental involvement in the development of children's reading skill: A five-year longitudinal study. *Child Development*, 73, 445–460.
- Sénéchal, M., LeFevre, J., Thomas, E. M., & Daley, K. E. (1998). Differential effects of home literacy experiences on the development of oral and written language. *Reading Research Quarterly*, 33, 96–116.
- Seymour, P. H. (2005). Early reading development in European orthographies. In M. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 296–315). Oxford, England: Blackwell.
- Seymour, P. H., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, 94, 143–174.
- Share, D. L. (1995). Phonological recoding and self teaching: Sine qua non of reading acquisition. *Cognition*, 55, 151–218.
- Spinelli, D., De Luca, M., Di Filippo, G., Mancini, M., Martelli, M., & Zoccolotti, P. (2005). Length effect in word naming in reading: Role of reading experience and reading deficit in Italian readers. *Developmental Neuropsychology*, 27, 217–235.
- Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, 24, 402–433.
- Sunseth, K., & Bowers, P. G. (2002). Rapid naming and phonemic awareness: Contributions to reading, spelling, and orthographic knowledge. *Scientific Studies of Reading*, 6, 401–429.



- Swanson, H. L., & Alexander, J. E. (1997). Cognitive processes as predictors of word recognition and reading comprehension in learning-disabled and skilled readers: Revisiting the specificity hypothesis. *Journal of Educational Psychology*, 89, 128–158.
- Swanson, H. L., & Howell, M. (2001). Working memory, short-term memory, and speech rate as predictors of children's reading performance at different ages. *Journal of Educational Psychology*, 93, 720–734.
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics*. Boston: Allyn and Bacon.
- Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of Word Reading Efficiency*. Austin, TX: PRO-ED.
- Torgesen, J. K., Wagner, R. K., Rashotte, C. A., Burgess, S., & Hecht, S. (1997). Contributions of phonological awareness and rapid automatized naming ability to growth of word-reading skills in second- to fifth-grade children. *Scientific Studies of Reading*, 1, 161–185.
- van den Bos, K. P. (1998). IQ, phonological awareness and continuous-naming speed related to Dutch poor decoding children's performance on two word identification tests. *Dyslexia*, 4, 73–89.
- Wagner, R. K., & Barker, T. A. (1994). The development of orthographic processing ability. In V. W. Berninger (Ed.), *The varieties of orthographic knowledge I: Theoretical and developmental issues* (pp. 243–276). Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. *Psychological Bulletin*, 101, 192–212.
- Wagner, R. K., Torgesen, J., & Rashotte, C., A. (1999). *CTOPP: Comprehensive Test of Phonological Processing*. Austin, TX: PRO-ED.
- Wagner, R. K., Torgesen, J., Rashotte, C. A., Hecht, S., Barker, T., Burgess, T., et al. (1997). Changing relations between phonological processing abilities and word-level reading as children develop from beginning to skilled readers: A 5-year longitudinal study. *Developmental Psychology*, 33, 468–479.
- Wechsler, D. (1992). *Wechsler intelligence scale for children* (3rd ed.). New York: Psychological Corporation.
- Wesseling, R., & Reitsma, P. (2000). The transient role of explicit phonological recoding for reading acquisition. *Reading and Writing: An Interdisciplinary Journal*, 13, 313–336.
- Wiederholt, J. L., & Bryant, B. R. (2001). *GORT 4: Gray Oral Reading Test*. Austin, TX: PRO-ED.
- Wimmer, H. (2006). Don't neglect reading fluency! *Developmental Science*, 9, 447–448.
- Wimmer, H., & Goswami, U. (1994). The influence of orthographic consistency on reading development: Word recognition, in English and German children. *Cognition*, 51, 91–103.
- Wimmer, H., Landerl, K., & Schneider, W. (1994). The role of rhyme awareness in learning to read a regular orthography. *British Journal of Developmental Psychology*, 12, 469–484.
- Wimmer, H., Mayringer, H., & Landerl, K. (2000). The double-deficit hypothesis and difficulties learning to read a regular orthography. *Journal of Educational Psychology*, 92, 668–680.
- Wolf, M., & Bowers, P. G. (1999). The double deficit hypothesis for the developmental dyslexias. *Journal of Educational Psychology*, 91, 415–438.
- Wolf, M., & Denckla, M. B. (2005). *Rapid Automatized Naming and Rapid Alternating Stimulus Tests (RAN/RAS)*. Austin, TX: PRO-ED.
- Woodcock, R. W. (1998). *Woodcock Reading Mastery Test—Revised Normative Update Examiner's Manual*. Circle Pines, MN: American Guidance Service.
- Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, 131, 3–29.
- Ziegler, J. C., & Goswami, U. (2006). Becoming literate in different languages: Similar problems, different solutions. *Developmental Science*, 9, 429–453.
- Ziegler, J. C., Perry, C., Jacobs, A. M., & Braun, M. (2001). Identical words are read differently in different languages. *Psychological Science*, 12, 379–384.
- Zoccolotti, P., De Luca, M., Di Pace, E., Gasperini, F., Judica, A., & Spinelli, D. (2005). Word length effect in early reading and in developmental dyslexia. *Brain and Language*, 93, 369–373.

Received February 16, 2007

Revision received November 15, 2007

Accepted December 16, 2007 ■

# Working Memory and Intelligence in Children: What Develops?

H. Lee Swanson  
University of California, Riverside

This study explored the contribution of the phonological and executive working memory (WM) systems to 205 (102 girls, 103 boys, 6 to 9 years old) elementary school children's fluid and crystallized intelligence. The results show that (a) a 3-factor structure (phonological short-term memory [STM], visual-spatial WM, and verbal WM) was comparable between age groups, (b) controlled attention and STM storage accounted for 67% of the age-related variance in WM, (c) effect sizes for direct paths from WM were substantially larger when predicting fluid intelligence than crystallized intelligence, and (d) the contribution of STM to intelligence was isolated to reading. The results suggest that the development of WM is distinct from STM, controlled attention plus storage accounted for age-related WM changes, and WM underlies age-related changes in both fluid and crystallized intelligence.

*Keywords:* working memory, short-term memory, phonological processing, executive processing

Working memory (WM) plays a critical role in learning during childhood, supported by studies demonstrating close links between measures of WM and measures of learning, achievement, and intelligence (see Gathercole, Lamont, & Alloway, 2006; Swanson & Siegel, 2001, for a review). WM has been referred to as a processing resource of limited capacity, involved in the preservation of information while simultaneously processing the same or other information (e.g., Baddeley, 2000; Engle, Tuholski, Laughlin, & Conway, 1999). Recent studies suggest that age-related performance of older children can be characterized as an ability to retain information in memory while simultaneously processing the same or other information when compared to younger children (e.g., Barrouillet & Camos, 2001; Conlin, Gathercole & Adams, 2005; Gavens & Barrouillet, 2004; Swanson, 1999). Age-related performance patterns are also observable in children on measures of fluid intelligence and skills related to measures of crystallized intelligence, such as reading and mathematics, as these areas of intelligence are considered to make large demands on WM capacities (e.g., Fry & Hale, 2000; Hitch, Towse, & Hutton, 2001; Siegel, 1994; Swanson, 1999). In these studies, WM has been viewed as distinct from short-term memory (STM). STM is viewed as involving situations where small amounts of material are held passively (i.e., minimal information is drawn from long-term memory to interpret the task, e.g., digits or word span tasks) and then reproduced in a sequential or untransformed fashion (e.g., Just & Carpenter, 1992).

## Distinction Between WM and STM

This study addressed three questions related to the development of WM in children. The first question addressed whether age-related changes in young children's WM are distinct from age-related changes in STM. An emerging consensus in the literature with adult participants is that WM and STM are distinct but highly related processes (e.g., see Heitz, Unsworth, & Engle, 2005, for a review). For example, Engle et al. (1999) found that STM and WM tasks loaded on two different factors (also see Cantor, Engle, & Hamilton, 1991; Conway, Cowan, Bunting, Theriault, & Minkoff, 2002, for similar findings). Although strong correlations emerged between the two factors ( $r = .70$ ), they found that a two-factor model fit the data better than a one-factor model. Further, they found that by statistically controlling the variance between WM and STM factors that the residual variance related to the WM factor was significantly correlated with measures of fluid intelligence (e.g., Raven Progressive Matrices Test [Raven, 1976] and the Cattell Culture Fair Test [Cattell, 1973]). That is, they found a strong link between the latent measures of WM, but not STM, to fluid intelligence. They interpreted their findings as suggesting that the residual variance related to the WM factor corresponded to controlled attention of the central executive system, and this system, in turn, was strongly correlated with measures of fluid intelligence.

The present study raises an important question as to whether STM and WM reflect two independent systems in children. One of the reasons for addressing this question is that STM and WM tasks have been found to load on the same factor in children. For example, Hutton and Towse (2001) found, via a principal-components analysis, that both WM and STM tasks loaded on the same factor for children 8 to 11 years old. Their results also showed that correlations related to WM and STM on measures of reading (.48 vs. .45), number skills (.33 vs. .38), and fluid intelligence (.35 vs. .36) were of the same magnitude (see Table 4), suggesting that WM and STM share the same construct (also see Cowan et al., 2003, for a similar finding). As stated by Hutton and Towse,

---

This article is based on a 3-year longitudinal study funded by U.S. Department of Education, Cognition and Student Learning (USDE R305H020055), Institute of Education Sciences awarded to H. Lee Swanson. I am indebted to Georgia Doukas, Diana Dowds, Rebecca Gregg, Krista Healy, Crystal Howard, James Lyons, Kelly Rosston, Olga Jerman, Xinhua Zheng, and Leilani Sáez for data collection and/or task development. Special appreciation is given to Bev Hedin, who directed and managed data collection and school schedules in the 2nd year of the project. This article does not necessarily reflect the views of the U.S. Department of Education or the school districts.

Correspondence concerning this article should be addressed to H. Lee Swanson, Graduate School of Education, University of California, Riverside, CA 92521. E-mail: lee.swanson@ucr.edu



"It appears that what holds for WM in adults may not be equally true for children, and vice versa. The study highlights the value of taking account of children's online processing during WM tasks, and in so doing suggests that WM and STM may, at least in some circumstances, be rather equivalent." (p. 392)

These findings, however, contrast with Gathercole, Pickering, Ambridge, and Wearing (2004), who investigated the structure of WM in children ages 4 to 15 years. Their results supported a basic modular structure of WM that was consistent with a tripartite adult-based model that includes phonological STM, executive processing, and the visual-spatial sketchpad.<sup>1</sup> However, it was unclear from Gathercole et al.'s (2004) results whether the processes that contributed to these structures were distinct. For example, the average factor correlation between measures of STM (phonological loop) and executive processing was .80 (see p. 186). In addition, the magnitude of this correlation increased across ages suggesting that the factor structure was less distinct for the older than younger participants. The authors indicated that the close association between the factors associated with the phonological STM and the executive component of WM was because "the central executive's identification was based on tasks that are constrained by phonological loop capacity" (p. 188). Given these findings, further work is necessary to determine whether children's STM is distinct from WM.

### Controlled Attention Versus Storage

The second question this study addresses is whether controlled attention and/or STM play the major role in age-related changes in children's WM span. In elaborating the distinction between STM and WM in children, Cowan (1995) emphasized the role of attentional processing in children's WM development. WM is described as a subset of items of information that are stored in STM and submitted to "controlled attentional" processing. *Controlled attention* is defined as the capacity to maintain and hold relevant information in the face of interference or distraction and is measured by partialing out the influence of STM from WM performance (Engle et al., 1999). The majority of measures used to tap some aspects of controlled attention involved inhibition (Rosen & Engle, 1997). Although controlled attention has been argued as a component of WM, the majority of research suggests that storage and speed underlie WM span. For example, McCabe and Hartman (2003) found that the locus of age effects on WM tasks for younger and older adults were due to an ability to store verbal material in STM and processing speed. Likewise, Bayliss, Jarrold, Baddeley, Gunn, and Leigh (2005) found with children (6 to 10 years of age) that processing speed and storage contributed to WM span. Thus, the role of controlled attention as underling children's WM span needs to be investigated further.

### Crystallized Versus Fluid intelligence

The third question addresses whether the influence of WM is primarily directed to fluid intelligence and/or crystallized intelligence. Fluid intelligence is typically assessed on standardized measures that tap reasoning, thinking, or the ability to acquire new knowledge, whereas crystallized intelligence is assessed on measures that tap what has been learned in a particular domain (e.g., reading and math; Cattell, 1971; Horn, 1980; Horn & Noll, 1997). *Fluid intelligence* is defined as the ability to reason (e.g., induce

abstract relations) under novel conditions, whereas *crystallized intelligence* is defined as academic achievement or cultural knowledge based on already learned knowledge (Cattell, 1971; also see Haavisto & Lehto, 2005, for an extensive review). Because verbal and numerical abilities are known to load on crystallized intelligence (Cattell, 1971), for this study I use reading and math tasks to reflect the verbal and numerical domains of crystallized intelligence, respectively.

I assumed that WM is essential for the mental activities basic to children's intelligence. For example, age-related changes in WM, with its capacity constraints, may provide an important foundation on intelligence measures related to children's reasoning (e.g., Fry & Hale, 2000), reading (e.g., Siegel, 1994), and mathematics (e.g., Geary, Hoard, & Hamson, 1999). The available evidence indicates that WM is an important part of the cognitive basis of intelligence (e.g., Case, 1992; Schweizer & Koch, 2002). However, because a part of WM distinguishes between STM and controlled attention, the question arises as to whether WM as an attentional mechanism or WM as storage (STM) is an especially important source of age-related differences in children's intelligence. Some authors emphasize the phonological storage aspect of WM (e.g., phonological loop) on measures of crystallized intelligence (i.e., reading; e.g., Shankweiler & Crain, 1986). Others suggest that efficiency of the executive system contributes to the relationship between WM and fluid intelligence (e.g., Case, 1992). Thus, the connecting mechanisms between WM and the two types of intelligence are unclear. It is possible that the executive system of WM (which manages a number of goals, representations, and procedures for problem solving, which requires controlled attention) acts as the critical WM factor for fluid intelligence tasks (e.g., Oberauer, Süß, Wilhelm, & Wittman, 2003), whereas the storage component of WM (accessing previously learned information) acts as the critical WM mechanism on crystallized tasks (e.g., see Hambrick & Engle, 2002, for the discussion on the role of domain-specific knowledge and WM). That is, WM is related to crystallized intelligence to the extent that WM reflects a resource that can be used to activate preexisting stored information and maintain that activation during the performance of some task. In addition, keeping and maintaining task-relevant information and inhibiting may be related to both fluid and crystallized intelligence. Thus, an investigation of the components (i.e., STM vs. controlled attention) of WM most related to children's intelligence is further warranted.

<sup>1</sup> In this model, Baddeley (1986; Baddeley & Logie, 1999) described WM as a limited capacity central executive system that interacts with a set of two passive store systems used for temporary storage of different classes of information: the speech-based phonological loop and the visual sketchpad. The phonological loop is responsible for the temporary storage of verbal information; items are held within a phonological store of limited duration, and the items are maintained within the store via the process of articulation. The visual sketchpad is responsible for the storage of visual-spatial information over brief periods and plays a key role in the generation and manipulation of mental images. Both storage systems are in direct contact with the central executive system. The central executive system is considered to be primarily responsible for coordinating activity within the cognitive system, but also devotes some of its resources to increasing the amount of information that can be held in the two subsystems (Baddeley & Logie, 1999). This model has been revised to include an episodic buffer (Baddeley, 2000), but support for the tripartite model has been found across various age groups (Gathercole et al., 2004).



Unfortunately, very little research has examined whether the processes that mediate the correlations between WM and crystallized intelligence measures are the same as those that mediate the correlations between WM and fluid intelligence measures. There are some studies suggesting that fluid and crystallized intelligence are more closely related in younger than older children because the school curriculum tends to standardize student's knowledge base (Schweizer & Koch, 2002). Thus, one may expect the magnitude of the correlations between WM and fluid intelligence and WM and crystallized intelligence to be more similar in younger than older children. There is some evidence from functional magnetic resonance imaging studies suggesting that development in fluid intelligence is related to the frontal lobe development (e.g., see Kane, 2005, for a review), whereas age-related changes on crystallized measures, such as the area of reading, are related to the maturation of temporal and parietal sections of the brain (e.g., Paulesu et al., 2001). Likewise, different components of WM have been related to the development of different regions of the brain. For example, STM (phonological loop) has been associated with the left temporal parietal region (Baddeley, 1986), whereas the executive system of WM is associated with the frontal lobes (T. D. Wagner & Smith, 2003; also see Kane, 2005, for a review). Thus, it seems reasonable from these studies that different processes as a function of age may mediate the influence of WM on different intelligence measures.

There is some evidence that the importance of WM in predicting intelligence varies with age. For example, the links between WM and fluid intelligence established in adults have not generalized to children. For example, Bayliss, Jarrold, Gunn, and Baddeley (2003) found with children (7 to 9 years old) that the residual variance between the two factors (WM and STM) was not linked to fluid intelligence (Raven Progressive Matrices Test) but was linked to measures of crystallized intelligence (reading and math). That is, in contrast to Engle et al.'s (1999) work with adult samples, no significant relationship emerged between children's complex span performance (WM) and performance on the Raven Progressive Matrices Test. Rather Bayliss et al. (2003) found that measures of processing efficiency (reaction time) and STM storage predicted performance on measures of reading and math, but not fluid intelligence.

### Phonological Loop Model

In summary, this study addresses questions as to whether WM and STM operate as distinct systems, whether controlled attention and/or storage mediate age-related changes in WM, and whether WM has more direct influence on one form of intelligence than another. To address these questions, I consider the contribution of the phonological and executive processes in accounting for age-related differences in WM and intelligence. Two models are considered. The first model hypothesizes that age-related differences in WM are strongly related to the phonological loop (STM).<sup>2</sup> The model assumes that because younger children's phonological system is more developed than their executive system when compared to older children, WM and STM are the same in younger children. Thus, WM and STM represent a single factor in young children. In contrast, STM and WM are distinct factors in older children. Indirect evidence for this assumption comes from studies linking the phonological system to a number of activities (e.g., language, reading) during early childhood years (e.g., Baddeley,

Gathercole, & Papagno, 1998), and therefore executive processing activities attributed to the WM system may develop much later in children than the phonological loop. There is some literature suggesting that tasks related to the phonological loop (i.e., STM) and executive system are not clearly distinguishable in young children (Gathercole et al., 2004). This lack of distinguishability may reflect the fact that the phonological system plays a more dominant role in WM performance than processes commonly attributed to the executive system (e.g., controlled attention).

I extend this research by investigating the processes that mediate the influence of WM on children's intelligence. A simple version of the above hypothesis states that younger children are less efficient at accessing and storing phonological information than older children and that such reduced processing on the children's part underlies children's lower performance on WM and intelligence measures. For example, one of the possible reasons WM span increases as a function of age is because older children store more information, which in turn is due to their ability to name items more rapidly than younger children (Henry & Millar, 1993; Hulme, Thomson, Muir, & Lawrence, 1984). Rapid naming is assumed to enhance the effectiveness of subvocal rehearsal processes and hence reduce the decay of memory items in the phonological store prior to output (e.g., Henry & Millar, 1993; Hulme et al., 1984; see Gathercole, 1998, for review). Naming speed has been interpreted as a measure of how quickly items can be encoded and rehearsed within the phonological loop (e.g., McDougall, Hulme, Ellis, & Monk, 1994). Another possible reason for age-related changes in WM is that older children are more aware of (or knowledgeable about) the sound structures in language than younger children. Several studies show that measures of verbal STM and phonological knowledge are strongly associated (e.g., see Alloway, Gathercole, Willis, & Adams, 2004, pp. 87–88, for review), and therefore greater phonological processing abilities are related to higher STM scores because they tap a common processing substrate (e.g., Griffiths & Snowling, 2002). There are clear expectations in this model. The influence of WM on measures of fluid and crystallized intelligence in children follows automatically with age-related improvements in the phonological loop.

### Executive Process Model

In contrast to the above model, the second model views age-related changes in the executive component of WM (i.e., controlled attention) as well as storage as underlying children's WM performance. That is, age-related changes in WM depend on both the phonological loop (STM) and the executive system. This model assumes there is unique variance related to both the executive and the phonological system of WM that contribute to children's intelligence. The executive component of interest is

<sup>2</sup> Previous research has already established links between the phonological system and STM (e.g., see Alloway, Gathercole, Willis, & Adams, 2004, for a review). For example, studies that have compared older and younger children have suggested that STM measures capture the utilization and/or operation of the phonological loop (see Gathercole, 1998; Gathercole & Baddeley, 1993, for a comprehensive review). The phonological loop has been referred to as STM (e.g., Baddeley, 1986; Baddeley, Lewis, & Vallar, 1984), because it involves two major components discussed in the STM literature: a speech-based phonological input store and a rehearsal process (see Baddeley, 1986, for review).



controlled attention.<sup>3</sup> Thus, in contrast to the aforementioned hypothesis that suggests phonological processes play a more dominant role in the mediating effects of WM on children's intelligence, this model assumes that executive processes (i.e., controlled attention) also play a major role in mediating both fluid and crystallized intelligence in children. Thus, there are clear expectations for this second model. The influence of WM on measures of fluid and crystallized intelligence in children follows automatically with improvements in controlled attention.

How might controlled attention contribute to age-related and individual differences in fluid and crystallized intelligence? To address this question, I administered two tasks that I assume are related to controlled attention: fluency and random generation. These tasks have been shown to tap two aspects of controlled attention processing: inhibition of competing responses during a controlled search for items (e.g., Rosen & Engle, 1997) and the filtering of rote or habitual responses (Towse, 1998). For example, the fluency task requires individuals to spontaneously generate words in response to a category cue (e.g., generate animal names) or specific letter cue (generate words that begin with the letter *b*). These tasks have been associated with the executive functions related to the controlled search for words (e.g., see Rende, Ramberger, & Miyake, 2002, for review). That is, participants are directed to activate needed information (animal names) while inhibiting the repetition of exemplars. In the random generation procedure, on the other hand, participants are asked to keep track of the number of times items have been generated and to inhibit well-known sequences such as "1, 2, 3, 4" or "a, b, c, d." Baddeley (1986) indicated that during random generation, the central executive acts as a filtering device, screening out automatically generated (and therefore nonrandom) responses. This task differs somewhat from the fluency measure because participants must inhibit rote or habitual responses (saying the letters of the alphabet in order) in order to quickly complete the task. Thus, during random generation the central executive acts as a rate-limited filtering device that filters out habitual responses (see Towse, 1998, for a review of this measure). Although not pure measures of inhibition, both tasks do require active suppression in order to perform the task efficiently.

### Purpose of Study

In summary, this study had two purposes. The first was to determine whether two independent memory systems related to STM and WM provide an adequate representation of transient memory in both younger and older children. The second purpose was to determine the processes that mediate age-related changes in the phonological loop and/or executive system and whether these same processes mediate the relationship between WM and intelligence. Because I assumed that age-related changes in WM may be due to age-related changes in children's phonological system, 8- and 9-year-old children's performance was compared with 6- and 7-year-old children's performance on measures of WM and STM. Engle et al. (1999) suggested that STM and WM might not be distinguishable in young children whose rehearsal processes are unstable. Further, several studies show weak to moderate correlations between memory span and articulation rates in children 7 years of age and below (see Gathercole, 1998, pp. 6–7, for review). This occurs because younger children's time-based strategies (i.e., rehearsal) are unstable (Gathercole, 1998). Thus, I se-

lected age levels where memory strategies are stable (8- and 9-year-olds) versus unstable (6- and 7-year olds) to better assess the processes that mediate the relationship between WM, STM, and intelligence.

### Predictions

Three predictions were tested in this study. First, a phonological loop and executive processing model provides an adequate representation of the data. If WM and STM are indeed distinct constructs in children, then an STM and a general WM factor should fit the data better than a one-factor solution. I also tested whether the separate phonological and executive factor structure holds for the two age groups.

Second, age-related changes in children's WM are mediated by both STM and controlled attention. Thus, both measures of STM and controlled attention predict age-related changes in WM.

Finally, WM predicts age-related increases in both fluid and crystallized intelligence. Likewise, WM draws from age-related increases in controlled attention and storage when predicting performance on fluid and crystallized measures of intelligence.

### Method

#### Participants

Two hundred and five children who were 6 to 9 years of age from a southern California public school district and private school district participated in this study. Final selection was related to parent approval for participation. Of the 205 participants, 102 were girls and 103 were boys. Children were divided into younger (ages 6 and 7,  $n = 98$ ) and older age groups (ages 8 and 9,  $n = 107$ ). Gender representation was not significantly different between the two age groups,  $\chi^2(1) = .0045, p > .05$ . Ethnic representation of the sample was 85 Anglo, 89 Hispanic, 18 African American, 9 Asian, and 4 other (e.g., Native American, Vietnamese). No significant differences emerged between the two age groups in terms of ethnicity,  $\chi^2(5) = 5.87, p > .05$ . The mean socioeconomic

<sup>3</sup> Some clarification between executive processing and controlled attention is necessary. The central executive is involved in the control and regulation of the WM system. Several cognitive activities have been assigned to the central executive (e.g., see Miyake, Friedman, Emerson, Witzki, & Howerter, 2000, for a review), such as controlling subsidiary memory systems, control of encoding and retrieval strategies, attention switching during manipulation of material held in the verbal and visual-spatial systems, suppressing irrelevant information, accessing information for long-term memory, and so on (e.g., Baddeley, 1996; Miyake et al., 2000; Oberauer et al., 2003). According to Baddeley and Logie (1999), the executive system coordinates the two subordinate systems (phonological loop and visual-spatial sketchpad), focusing and switching attention, in addition to activating representations within long-term memory. Thus, several activities could be assigned to the central executive. Some of these activities have been reduced to three functions: (a) updating and monitoring of working memory representations, (b) inhibition of dominant or prepotent responses, and (c) shifting between mental sets (Miyake et al., 2000). In this study, I focus on one aspect of the executive system: controlled attention. Controlled attention is operationally defined as the residual variance left in WM when STM has been partialled from the analysis (Engle et al., 1999). I assumed that this residual variance (i.e., controlled processing) is reflected in activities that involve the inhibition of competing information from the targeted information.

status of the sample was primarily middle to upper middle class on the basis of parent education or occupation. Means and standard deviations for the selection variables used in this study are shown in Table 1.

### Tasks and Materials

The battery of group and individually administered tasks is described below. Experimental tasks are described in more detail than published and norm referenced tasks. All experimental tasks included two practice trials.

### Measures of Fluid and Crystallized Intelligence

**Fluid intelligence.** Fluid intelligence was assessed by the Raven Progressive Matrices Test (Raven, 1976). Children were given a booklet with patterns displayed on each page, each pattern revealing a missing piece. For each pattern, four possible replacement pattern pieces were displayed. Children were required to circle the replacement piece that best completed the patterns. After

the introduction of the first matrix, children completed their booklets at their own pace. Patterns progressively increased in difficulty. The dependent measure (range = 0–36) was the number of problems solved correctly.

**Arithmetic computation.** The Arithmetic subtest from the Wide Range Achievement Test—Third Edition (WRAT-III; Wilkinson, 1993) was administered. The dependent measure was the number of problems correct, which yielded a standard score.

**Reading.** Reading was assessed by the Word Recognition subtest of the WRAT-III. The task provided a list of words of increasing difficulty. The child's task was to read the words until 10 errors occurred. The dependent measure was the number of words read correctly.

### Phonological Knowledge Measures

Because the phonological measures are commonly used and derived from published standardized measures (i.e., Woodcock Reading Mastery Test—Revised [Woodcock, 1998], Wechsler In-

Table 1  
Means and Standard Deviations of Dependent Measures

Measure	Younger ( <i>n</i> = 98)			Older ( <i>n</i> = 107)			<i>F</i> (1, 203)	$\eta^2$
	<i>M</i>	<i>SD</i>	KR <sub>20</sub>	<i>M</i>	<i>SD</i>	KR <sub>20</sub>		
Age	6.57	.29		8.73	.36			
Fluid Intelligence								
Raven (stand.)	106.00	14.33	.85	104.27	13.70	.90	0.78	.01
Raven (raw)	19.89	5.41	.85	26.02	5.53	.82	64.07**	.24
Mathematics								
WRAT-III (stand.)	114.84	9.60	.91	112.17	14.54	.92	2.61	.01
WRAT-III (raw)	19.05	1.68	.80	28.72	3.77	.70	544.58**	.73
Reading-Word Recognition								
WRAT-III (stand.)	111.04	16.81	.90	102.91	13.38	.87	14.77**	.07
WRAT-III (raw)	23.02	5.22	.78	31.93	4.85	.65	160.26**	.44
Phonological Knowledge								
1. Elision	7.28	4.98	.92	11.80	5.71	.92	36.09**	.15
2. Pseudowords	28.41	16.96	.96	61.30	13.70	.91	234.84**	.53
Speed								
Rapid letter naming	62.26	19.25	.92	40.48	10.26	.80	104.47**	.34
Rapid number naming	52.86	14.25	.89	35.43	9.43	.81	108.18**	.35
Controlled Attention								
Fluency								
Category-Animal	11.15	4.54	.77	14.71	4.34	.67	33.48**	.14
Letter-B	6.09	3.36	.72	7.85	3.24	.72	14.61**	.07
Random Generation								
Random-letter	6.20	2.45	.51	8.11	3.27	.54	21.95**	.10
Sequence-letter	11.24	5.12	.84	19.30	6.55	.80	95.05**	.32
Random-number	4.14	2.98	.74	6.12	2.97	.73	22.52**	.10
Sequence-number	17.98	6.52	.90	32.57	7.38	.81	222.76**	.52
Difference score-letter	5.81	4.29		14.05	6.20		50.29**	.20
Difference score-number	11.47	6.71		26.44	7.14		174.32**	.46
Short-Term Memory								
Digit span forward	2.56	.97	.67	3.28	1.09	.83	24.08**	.09
Pseudoword span	1.94	1.10	.85	2.41	.94	.84	9.27*	.04
Word span	3.54	.81	.71	3.96	.69	.61	12.28*	.05
Working Memory								
Mapping and Directions	1.68	.87	.59	2.45	1.50	.61	19.30**	.09
Visual Matrix	3.01	3.23	.81	5.73	4.05	.85	28.04**	.12
Listening Span	.86	1.31	.78	2.07	1.61	.77	34.04**	.14
Conceptual Span	1.92	.87	.77	2.42	.89	.78	15.85**	.07

Note. KR<sub>20</sub> = Kuder-Richardson Formula 20; stand = standard score from normed referenced test; Raven = Raven Color Progressive Matrices Test; WRAT-III = Wide Range Achievement Test—Third Edition.

\* *p* < .05. \*\* *p* < .01.



telligence Scale for Children—Third Edition [Psychological Corporation, 1991], Comprehensive Test of Phonological Processing (CTOPP; R. Wagner, Torgesen, & Rashotte, 2000), I only briefly describe these tasks.

*Pseudoword reading task.* The Pseudoword subtest was administered from the Test of Word Reading Efficiency (R. Wagner & Torgesen, 1999). The subtest required oral reading of a list of 120 real words or pseudowords of increasing difficulty. Children were given 45 s to say aloud as many words as possible from a list of nonwords. The nonwords followed regular spelling patterns, requiring children to quickly decipher pronunciations based on their existing knowledge of phonology. The dependent measures for both subtests were the number of words said correctly in 45 s.

*Elision or phonological deletion.* The Elision subtest from the CTOPP (R. Wagner et al., 2000) was administered. The Elision subtest measures the ability to parse and synthesize phonemes. The child was asked to say a word and to say what word is left if part of the word is deleted. There are four practice items and 15 test items. The dependent measure was the number of items said correctly.

### Naming Speed

*Digit naming speed.* The administration procedures followed those specified in the manual of the CTOPP (R. Wagner et al., 2000). For this task, the examiner presented participants with an array of 36 digits. Participants were required to name the digits as quickly as possible for each of two stimulus arrays containing 36 items, for a total of 72 items. The task administrator used a stopwatch to time participants on speed of naming. The dependent measure was the total time to name both arrays of numbers. The correlation between array Form A and B was .91.

*Letter naming speed.* The administration procedures followed those specified in the CTOPP (R. Wagner et al., 2000), including the presentation of practice trials. The manual reports correlations between parallel forms ranging from .80 to .93. For this task, the examiner presented participants with an array of 36 letters. Participants were required to name the letters as quickly as possible for each of two stimulus arrays containing 36 letters, for a total of 72 letters. The task administrator used a stopwatch to time participants on speed of naming. The dependent measure was the total time to name both arrays of letters. The correlation between array Form A and B was .90.

### Inhibition

Verbal fluency tasks required individuals to spontaneously generate words for 60 s, in response to a category cue (e.g., generate animal names) or specific letter cue (generate words that begin with the letter *b*). These tasks have been associated with frontal lobe function, and individual differences have been linked to executive functions (e.g., controlled search for words; e.g., see Rende, Ramsberger, & Miyake, 2002, for review). In this study, fluency tasks were presumed to reflect monitoring of responses to avoid repetitions during the generation of multiple response alternatives. Two verbal fluency tasks were administered and are described below.<sup>4</sup>

*Semantic fluency.* The experimental measure was adapted from Harrison, Buxton, Husain, and Wise (2000). Children were given 60 s to generate as many names of animals as possible.

Children were told, "I want to see how many animals you can name. Try not to repeat yourself. Don't say pet names. Keep going until I tell you to 'stop.' Ready, begin." Repetitions were deleted from the analysis. The dependent measure was the number of words correctly stated within 60 s.

*Phonological fluency.* This experimental measure was also adapted from Harrison et al. (2000). Children were given 60 s to generate as many words as possible beginning with the letter *b*. Children were told,

I want to see how many words you can say that begin with a certain letter. Do not say proper nouns or numbers or the same word with different endings and try not to repeat yourself. Keep naming words that start with the letter until I say "stop." Speak clearly and loudly enough so that I can hear the word you are saying. Do you understand? The letter is *b*. Begin.

Repetitions, proper name errors, as well as contravention of the stem repetition were deleted from the analysis. The dependent measure was the number of words correctly stated in 60 s.

*Sequential and random generation of letters and numbers.* The random generation task has been well articulated in the literature (e.g., Baddeley, 1996; Towse, 1998). The task measures controlled attention because participants are required to actively monitor candidate responses and suppress responses that would lead to well-learned sequences, such as 1-2-3-4 or a-b-c-d (Baddeley, 1996). Because this task has been primarily used with adult samples who have quicker access to letters and numbers, it was modified for the age group in this study. Each child was asked to write as quickly as possible numbers (or letters) in sequential order to establish a baseline of their ability to quickly access learned information. Children were then asked to quickly write numbers (or letters) in a random nonsystematic order. For example, for the number section, children were first asked to write numbers from 0 to 9 in order (i.e., 1, 2, 3, 4) and repeatedly as quickly as possible in a 30-s period. They were then asked to write numbers as quickly as possible "out of order" within a 30-s period. Two scores were calculated: speeded order retrieval and nonorder generation. For the speeded retrieval score, the total number of numbers or letters written in correct order was scored. The measure of random generation was calculated as the number of letters or numbers out of sequential order. As a measure of inhibition, random generation scores (corrected for redundancy and order) were subtracted from the speed retrieval (nonrandom generation) score. The scores were combined because a preliminary analysis of the data showed a tradeoff between speed and accuracy. Thus, it was assumed that high difference scores reflected a tradeoff between accuracy and speed.

### STM Measures

Three measures of STM were administered: Forward Digit Span, Word Span, and Pseudoword Span. The Forward Digit Span

<sup>4</sup> I assumed that the switching between the semantic retrieval aspect of the fluency task taps attentional control and the ability to inhibit previous responses (e.g., Rende et al., 2002; Rosen & Engle, 1997). As indicated by one reviewer, however, multiple processes can contribute to fluency tasks besides inhibition, including strategic initiation, sustain effort in conducting searches, and associative-semantic retrieval processes. Thus, I agree with the reviewer that I do not have a pure measure of inhibition.



subtest from the Wechsler Intelligence Scale for Children—Third Edition was administered. The Forward Digit Span task required participants to recall and repeat in order sets of digits that were spoken by the examiner and that increased in number. The Word Span and Pseudoword Span tasks were presented in the same manner as the Forward Digit Span measure. The Word Span task was previously used by Swanson, Ashbaker, and Lee (1996). The word stimuli are one- to two-syllable high-frequency words. Children are read lists of common but unrelated nouns and then are asked to recall the words. Word lists gradually increased in set size, from a minimum of two words to a maximum of eight. The Pseudoword Span task (Swanson & Berninger, 1995) used strings of nonsense words (one syllable long), which are presented one at a time in sets of 2–6 nonwords (e.g., *zeb, vab; sme, pru, tri*). The dependent measure for all STM measures was the highest set of items retrieved in the correct serial order (scores ranged from 0 to 7).

### WM Measures

The WM tasks in this study required children to hold increasingly complex information in memory while responding to a question about the task. The questions served as distracters to item recall because they reflected the recognition of targeted and closely related nontargeted items. A question was asked for each set of items, and the tasks were discontinued if the question was answered incorrectly or if all items within a set could not be remembered. Thus, WM span reflected a holding of some information (item storage) during the processing of other information (correct responses to questions). Consistent with Daneman and Carpenter's (1980) seminal WM measure, the processing of information was assessed by asking participants simple questions about the to-be-remembered material, whereas storage was assessed by accuracy of item retrieval. The question required a simple recognition of new and old information and was analogous to the yes/no response feature of Daneman and Carpenter's task. It is important to note, however, that in my tasks the difficulty of the processing question remained constant within task conditions, thereby allowing the source of individual differences to reflect increased storage demands. Further, the questions focused on the discrimination of items (old and new information) rather than deeper levels of processing such as mathematical computations (e.g., Towse, Hitch, & Hutton, 1998). A previous study with a different sample established the reliability and the construct validity of the measures with the Daneman and Carpenter measure (Swanson, 1996, 1999). For this study, three WM tasks were selected from a standardized battery of 11 WM tasks because of their high construct validity and reliability (see Swanson, 1992). The complete description of administration and scoring of the tasks are reported in Swanson (1995). A children's adaptation of the Daneman and Carpenter measure (Swanson, 1992; Swanson & Ashbaker, 2000) was also administered. Each WM task was scored by the highest set of items achieved when the process question was correct. However, partial credit (.50) was given if the process question was correct but not all items in the array were recalled perfectly. For example, if a child correctly recalled Set 1 but missed recalling perfectly the next highest set (Set 2) but got the process question correct, his or her WM score was 1.5. For all WM and STM tasks, children were presented two practice trials. After successful completion of the

practice trials, children were provided sets of items that increased in difficulty. Task descriptions follow.

*Mapping and directions.* This task required the child to remember a sequence of directions on a map (Swanson, 1992, 1995). The experimenter presented a street map with dots connected by lines; arrows illustrated the direction a bicycle would go to follow this route through the city. The dots represented stoplights, while lines and arrows mapped the route through the city. The child was allowed 10 s to study the map. After the map was removed, the child was asked a process question. The process question asked the children to indicate whether a stop line was in a particular column. The child was then presented a blank matrix on which to draw the street directions (lines and arrows) and stop lights (dots). Difficulty ranged on this subtest from 4 dots to 19 dots. The dependent measure was the highest set of a correctly drawn map (range = 0–9) in which the process question was answered correctly.

*Visual matrix task.* The purpose of this task was to assess the ability of participants to remember visual sequences within a matrix (Swanson, 1992, 1995). In contrast to the standardization procedures (Swanson, 1995), the visual matrix task was administered in small groups. An overhead projector was used to display stimuli to groups of children instead of individually by use of the examiner's manual. This change in format required students to circle their answer to the process question, rather than to verbally respond. Otherwise, the task was administered as per the manual instructions. Participants were presented a series of dots in a matrix and were allowed 5 s to study the matrix. After the matrix was removed, the child was asked a process question. The process question asked the children to indicate whether there were dots in a particular column. To ensure the understanding of columns prior to testing, participants are shown the first column location and practice finding it on blank matrices. In addition, for each test item the experimenter pointed to the first column on a blank matrix (a grid with no dots) as a reminder of first column location. After answering the discriminating question (by circling *y* for yes or *n* for no), students were asked to draw the dots they remembered seeing in the corresponding boxes of their blank matrix response booklets. The task difficulty ranged from a matrix of 4 squares and 2 dots to a matrix of 45 squares and 12 dots. The dependent measure was the highest set recalled correctly (range of 0 to 11) in which the process question was answered correctly.

*Conceptual Span.* The purpose of this task was to assess the participant's ability to organize sequences of words into abstract categories (Swanson, 1992, 1995). The participant was presented a set of words (one every 2 s), asked a discrimination question, and then asked to recall the words that go together. For example, a set might include the following words: *shirt, saw, pants, hammer, shoes, nails*. Participants were directed to retrieve the words that go together (i.e., *shirt, pants, and shoes; saw, hammer, and nails*). The process or discrimination question asked the child to choose whether a targeted word or nontargeted word was in the list of words. Thus, the task required participants to transform information encoded serially into categories during the retrieval phase. The range of set difficulty was two categories of two words to five categories of four words. The dependent measure was the highest set recalled correctly (range of 0 to 8) in which the process question was answered correctly.

*Listening Sentence Span.* The children's adaptation (Swanson, 1992) of Daneman and Carpenter's (1980) Sentence Span task was administered. The construction of, and pattern of results associated



with, the two measures are comparable. The only difference was that each sentence was read to the child with a 5-s pause that indicated the end of a sentence. The original Sentence Span measure was used with university students, whereas the current measure uses simpler sentences and reading vocabulary. WM capacity was defined as the largest group of ending words recalled. The mean sentence reading level (based on a Fleisher reading index) was approximately 3.8. The dependent measure was the highest set recalled correctly (range of 0 to 8) in which the process question was answered correctly.

### Reliability

The reliability of the measures was computed for this sample. Because the subsequent analysis relied on raw scores for analysis, the reliability of all raw score measures were calculated. Further, because reliability is inflated across a large age range, I calculated the reliabilities within each age group. As shown in Table 1, Kuder–Richardson Formula 20 scores varied, with the majority of coefficients hovering around .70. Further, the majority of reliability coefficients are within an acceptable range for basic research (around .70; see Nunnally & Bernstein, 1994, pp. 264–265, for discussion).

### Procedure

Three doctoral-level graduate students trained in test administration tested all participants in their schools. Two sessions of approximately 45–60 min each were required for small group test administration and one session of 45–60 min for individual test administration. During the group testing session, data were obtained from the Raven Progressive Matrices Test and the WRAT-III. The remaining tasks were administered individually. Test administration was counterbalanced to control for order effects. Task order was random across participants within each test administrator.

## Results

The means and standard deviations for measures of fluid intelligence, crystallized intelligence (reading, arithmetic calculation), phonological knowledge, naming speed, sequential and random generation, fluency, STM, and WM are shown in Table 1. Table 1 also reports the reliability of each measure and the univariates and effect sizes ( $\eta^2$ ) for each comparison. An  $\eta^2$  of .13, .05, and .02 corresponded to Cohen's  $d$  of .80, .50, and .20, respectively. The analyses and results were divided into two sections. The first section focused on age differences. The second section focused on correlations between STM and WM in the complete sample. This approach allowed us to study the entire range of scores in WM. I used a path analysis to test whether controlled attention (inhibition in this case) and/or storage (STM in this case) predicted age-related changes in WM. Hierarchical regression models were computed to determine the mediating role of WM on age-related changes in fluid and crystallized intelligence.

### Age Comparisons

A series of multivariate analyses of covariances were conducted to examine age-related differences in raw scores for accuracy or span for (a) fluid and crystallized intelligence (Raven,

arithmetic calculation, word recognition), (b) STM/WM, (c) naming speed/random generation, and (d) and phonological knowledge/fluency. The univariates are shown in Table 1. Alpha was set to .05 unless otherwise specified.

A significant multivariate main effect emerged for age on measures of crystallized and fluid intelligence, Wilks's  $\Lambda = .27$ ,  $F(3, 201) = 180.66$ ,  $p < .0001$ ; STM/WM, Wilks's  $\Lambda = .70$ ,  $F(7, 197) = 12.02$ ,  $p < .0001$ ; naming speed/phonological knowledge, Wilks's  $\Lambda = .59$ ,  $F(4, 200) = 33.80$ ,  $p < .0001$ ; and random generation/fluency, Wilks's  $\Lambda = .49$ ,  $F(4, 200) = 50.75$ ,  $p < .0001$ . Also shown are the significant effects related to difference scores from the random generation task. As expected, all univariates were significantly in favor of older children when compared to younger children. In general, the results clearly show that older children outperformed younger children across measures of fluid and crystallized intelligence, naming speed/controlled attention, STM/WM, and phonological knowledge.

### Correlation/Confirmatory Factor Analysis

Prior to my analysis of various processes assumed to underlie age-related changes in STM and WM performance, I investigated the correlation structure of WM and STM tasks for the total sample. The intercorrelations among the memory measures as well as process and criterion measures of the total sample ( $N = 205$ ) are shown in Table 2.<sup>5</sup> One finding worth noting was that the magnitude of the correlations within verbal WM measures (listening span and conceptual span,  $r = .62$ ) and visual spatial measures (mapping and visual matrix,  $r = .41$ ) was higher than between verbal and visual spatial measures ( $r$ s ranged from .11 to .30). This finding suggests that the memory tasks captured a diversity of memory operations.

*Confirmatory analysis.* Our hypothesis was that the structure of the memory measures, as reflected in Baddeley and Logie (1999), reflected three factors: phonological loop, visual–spatial sketchpad, and executive processing. In order to test this hypothesis, I first ran a confirmatory factor analysis with all the memory measures loading one factor. I then ran a second model with the WM tasks (mapping/direction, visual matrix, listening span, conceptual span) loading on one factor and the three STM tasks (forward digit, pseudoword, word span) on a separate factor. I then ran a third model with the two visual–spatial WM tasks (mapping/direction, visual matrix) on one factor, the two verbal WM tasks on a second factor (listening span, conceptual span), and the three STM tasks on a third factor. The fit statistics for the three models are shown in Table 3. Data were analyzed using the EQS 6.1

<sup>5</sup> Prior to testing my hypothesis, normalcy of the data was considered. As shown in Table 2, all memory measures meet standard criteria for univariate normality (Kline, 1998) with skewness for all measures less than 3 and kurtosis less than 4. Univariate outliers were defined as cases more than 3.5 standard deviations from the means. Multivariate outliers were examined by calculating Mahalanobis'  $d^2$ . None of the cases related to memory performance were deemed outliers. The data were also checked for multicollinearity. Except for rapid naming speed between letters and numbers, none of the zero-order correlations were above .80. One means to handle multicollinearity (as with the rapid naming measures) is to combine the scores into a latent measure, which occurs in the subsequent analysis. Overall, the zero-order correlations indicated that multicollinearity was not a problem with the memory data at the measurement level (Kline, 1998).

Table 2  
Intercorrelation Among Memory, Cognitive, and Intelligence Measures (N = 205)

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1. Age	—																		
Phonological Knowledge																			
2. Elision	0.54	—																	
3. Pseudowords	0.35	0.60	—																
Rapid Naming																			
4. Rapid Naming-Letter	-0.55	-0.67	-0.47	—															
5. Rapid Naming-Digit	-0.56	-0.73	-0.47	-0.92	—														
Fluency																			
6. Letter	0.22	0.35	0.28	-0.35	-0.37	—													
7. Category	0.35	0.42	0.41	-0.42	-0.45	0.37	—												
Random Generation																			
8. Letters	0.41	0.48	0.37	-0.45	-0.51	0.25	0.31	—											
9. Numbers	0.64	0.57	0.34	-0.60	-0.64	0.31	0.28	0.55	—										
Short-term memory																			
10. Forward Digits	0.29	0.31	0.19	-0.35	-0.32	0.20	0.12	0.23	0.55	—									
11. Pseudoword Span	0.19	0.31	0.31	-0.33	-0.35	0.18	0.30	0.22	0.24	0.18	—								
12. Word Span	0.22	0.26	0.28	-0.30	-0.34	0.18	0.22	0.30	0.23	0.32	0.35	—							
Working memory																			
13. Map Recall	.30	.25	.26	-.19	-.18	.17	.23	.19	.29	.02	.09	.12	—						
14. Visual Matrix	0.36	0.26	0.12	-0.31	-0.29	0.15	0.10	0.18	0.29	0.12	.05	.04	.41	—					
15. Sentence Span	0.37	0.20	0.38	-0.34	-0.39	0.23	0.40	0.38	0.19	0.17	0.30	0.35	0.30	0.28	—				
16. Conceptual Span.	0.27	0.43	0.25	-0.33	-0.32	0.24	0.30	0.27	0.18	0.24	0.29	0.24	0.11	0.22	0.62	—			
Fluid/Crystallized Intelligence																			
17. Raven	0.51	0.33	0.46	-0.43	-0.45	0.30	0.37	0.24	0.39	0.24	0.20	0.20	0.32	0.27	0.42	0.25	—		
18. Math	0.82	0.56	0.52	-0.64	-0.65	0.29	0.44	0.53	0.27	0.34	0.24	0.26	0.35	0.37	0.49	0.35	0.59	—	
19. Reading	0.61	0.70	0.61	-0.70	-0.74	0.22	0.43	0.47	0.54	0.35	0.30	0.27	0.33	0.31	0.42	0.30	0.57	0.62	—
M	92.38	21.94	9.96	50.89	43.77	7.01	13.02	8.77	20.52	2.94	2.19	3.76	2.08	4.43	1.49	2.18	23.09	24.10	27.67
SD	13.92	13.92	5.82	18.70	14.80	3.41	4.78	6.33	9.13	1.09	1.04	.78	1.30	3.92	1.59	.91	6.26	5.67	6.72
Skewness	-.01	.24	.32	1.61	1.18	.44	0.09	.88	.02	.36	-.98	-.66	1.33	.47	.56	.08	-.01	.24	-.25
Kurtosis	-1.64	-.68	-1.02	3.42	2.03	.08	.13	1.15	-.63	.59	.41	1.87	1.68	.84	-1.25	1.87	-.90	-1.44	-.63

Note. Raven = Raven Color Progressive Matrices Test.



Table 3  
*Fit Statistics for Confirmatory Factor Analysis and Structural Equation Models*

Model	df	$\chi^2$	<i>p</i>	NNFI	CFI	RMSEA
Confirmatory analysis						
One factor	13	68.256	.000	.742	.773	.14
Two factor	11	44.06	.00001	.833	.864	.12
Three factor	8	17.136	.028	.901	.962	.075
Second order	7	17.639	.013	.901	.956	.086
Multisample						
Null model	42	203.373	.0001			
Factor invariance (three factor)	33	47.89	.045	.88	.91	.045
Partial invariance <sup>a</sup> (three factor)	32	41.54	.12	.92	.94	.038
Age-related model predicting working memory						
Null model	120	1556.517	< .00001			
Theoretical model	81	153.55	< .001	.93	.95	.06
Predictive model for intelligence						
Null model	136	1924.31	< .0001			
Theoretical model	92	228.28	< .0001	.89	.92	.09
Revised model	90	201.69	< .001	.91	.94	.08

Note. NNFI = Bentler-Bonett nonnormed fit index; CFI = comparative fit index; RMSEA = root-mean-square error of approximation.

<sup>a</sup> Partial invariance removed the constraint that Visual Matrix loaded on the same factor for both age groups.

(Bentler, 2006), and the models tested were covariance structure models.

The chi-square statistic included in Table 3 provided a test of the null hypothesis that the model fits the data. If the model provided a good fit, the chi-square will be relatively small and corresponding *p* values will be relatively large. However, the chi-square is found to be significant even if the model is a good fit (Bentler & Wu, 1995). Therefore, researchers seek a small chi-square rather than a nonsignificant chi-square. Several studies recommend the nonnormed fit index (NNFI; Bentler & Bonnet, 1980) and the comparative fit index (CFI; Bentler & Wu, 1995) as overall goodness-of-fit indices. These indices are less likely to produce biased estimates in small samples. Values at .90 and over on the NNFI and CFI indicated an acceptable fit. The root-mean-square error of approximation measured the average residual correlation (Jöreskog & Sörbom, 1984). Small values (e.g., .10 or less) reflected a good fit.

As shown in Table 3, all the fit indices indicated that the three-factor model was a better fit to the data than the one- or two-factor model. A chi-square difference test indicated that the three-factor model did indeed fit the data significantly better than the two-factor model,  $\Delta\chi^2(1) = 64.43, p < .001$ . The three-factor model is shown in Figure 1 (top panel). Although the results provide an excellent fit to a three-factor model, it makes theoretical sense that the two factors (verbal WM and visual-spatial WM) share variance with a second-order factor. That is, both the verbal and visual-spatial WM measures were assumed to be monitored by a common source (executive system) because the tasks involve both simultaneous processing and storage. Thus, I computed a second-order factor to reflect cross-domain processing among the visual-spatial and verbal WM measures. The parameter estimates for the second-order model are shown in Figure 1 (bottom panel), and the indices for the model fit are reported in Table 3. Not surprisingly, the chi-square statistics for the

second-order factor model are no different than those presented for the three-factor model. This is because the two models were statistically indistinguishable but represent different hypotheses about the data. Although the chi-square statistic was significant, the model's  $\chi^2/df$  (2.52) was less than 3, which indicated an adequate fit to the data (Kline, 1998).

*Multiple group comparison.* An important question in this study was whether the covariance matrix fit both the younger and older children. A multisample analysis using structural equation modeling (Bentler, 2006) was computed on the covariance matrix for younger and older children. This analysis was done to answer the question as to whether WM and STM were distinct processes in young children. I tested whether a three-factor correlated model in which the STM measures (Forward digit, word span, pseudoword) loaded on one factor, visual-spatial WM measures (mapping/direction, visual matrix) on a second factor, and verbal WM (sentence span, conceptual span) loaded on a third factor held for both age groups. The theoretical model included cross-group constraints (Bentler, 2006). That is, the factor loadings and factor covariance were constrained to be equal across the two samples. Meredith (1993) considered this level of measurement configural invariance where the number of factors and pattern of factor loadings are considered to be identical across groups. The statistical fit for the two groups on the theoretical model is shown in the middle of Table 3. The model provides an excellent fit indicating that both age groups were comparable in terms of factor structure and loadings. However, follow-up stepwise Lagrange multiplier tests indicated that the constraint that accounted for most of the decrease in fit was the loading of visual matrix task on the visual-spatial factor for younger children. Thus, a revised model was tested where the visual matrix was not constrained to be equal across the two age groups. A comparison between the revised model and the theoretical model was significant (47.89–41.54),  $\Delta\chi^2(1) = 6.35, p < .05$ .

This finding indicated that the revised model was also a good fit to the data. Taken together, however, these results show that the two age groups did not differ in the structure of memory.

### *Path Analysis With WM as a Criterion (Endogenous) Variable*

**Path analysis.** The next analysis determined those variables that best accounted for age-related changes in WM. This analysis was done to answer the question as to whether age-related changes in children's WM are mediated by both STM and controlled attention. To this end, data were again analyzed using the EQS 6.1 (Bentler, 2006) and covariance structure models. Standard deviations and intercorrelations for all the variables are shown in Table 2. I report the null model and the theoretical model.

A diagram of the theoretical model tested is shown in Figure 2. To simplify the figure, the residuals are not shown. As shown, the theoretical model shows that age-related increases in performance on inhibition measures (fluency and random generation) influenced the second-order factor (executive processing). In addition, phonological knowledge and naming speed had direct paths to STM. I also included a path from age to STM because recent findings indicated that the development of storage ability (e.g., Bayliss et al., 2003, 2005) contributed unique variance to children's WM. To look at the separate contribution of the inhibition measures to WM, a path was established from fluency to random generation. I assumed that the two measures reflect different activities related to controlled attention: One was related to accessing knowledge and controlling for repetitions (fluency), whereas the other suppressed well-learned sequences (random generation). Likewise, a direct path was included from naming speed to phonological knowledge to partial the effects of each process on STM storage.

As shown in Table 3, the theoretical model was a good fit to the data. All paths were significant, and the results were consistent with my hypothesis suggesting that growth in controlled attention (fluency and random generation) is related to WM, and growth in speed and phonological knowledge is related to age-related changes in the phonological loop (STM). The theoretical model in Figure 2 shows a significant direct effect for fluency (standardized coefficient = .46), random generation (standardized coefficient = .27), and STM (standardized coefficient = .26) in predicting WM span. As shown, the disturbance term of the latent variable was .57. From the disturbance term (variance unaccounted for in the latent measure;  $D$ ) I can calculate  $R^2$  ( $R_j^2 = 1 - D_j^2$ ) or the proportion of variance accounted for in the direct path,  $j$ . The direct paths to WM accounted for 67% of the variance and therefore only 33% of the variance was unaccounted for. For the STM latent measure, significant direct paths were found for age (standardized path coefficient = .33), naming speed (standardized path coefficient = -.29), and phonological knowledge (standardized path coefficient = .06), which combined accounted for 41% of the variance. An interpretation of these standardized path coefficients can follow Cohen's (1988) recommendations where an absolute value of .10 represents a small effect, .30 represents a medium effect, and  $> .50$  represents a large effect. Thus, a small effect was found for phonological knowledge in predicting STM. A Wald test was computed to determine whether some of the paths could be deleted without changing model fit. The path from phonological knowledge to STM could be dropped. However, this path was

maintained because the path was significant in the theoretical model. I also investigated indirect effects on STM and WM.

The only significant mediating variable for WM was increases in the manifest variable of age (unstandardized path = .29,  $SE = .12$ , standardized = .76,  $t = 2.41$ ,  $p < .05$ ). No significant indirect paths were found for STM (all  $ps > .05$ ).

In summary, there were three important findings related to this analysis. First, the results show that activities related to controlled attention (fluency and random generation) contributed to age-related changes in WM span. Although previous studies have shown that WM span is related to speed and STM storage (Bayliss et al., 2005; Fry & Hale, 2000), the current study also shows that activities related to controlled attention contributed (in this case inhibition) important variance to WM span. Second, based on the magnitude of the standardized path coefficients, the largest effect size in terms of direct paths to WM was fluency. Finally, I found that the direction of the standardized path coefficients from speed to phonological knowledge and fluency to random generation was significant. These findings suggest that rapid naming has a direct effect on phonological knowledge and fluency has a direct effect on random generation.

### *Path Analysis With Fluid and Crystallized Intelligence as Criterion (Endogeneous) Measures*

For my next set of analyses, I determined whether the age-related paths in STM, fluency, and random generation to WM were maintained when WM predicted performance on measures of fluid and crystallized intelligence. I address the question as to whether WM predicts age-related increases in both fluid and crystallized intelligence. I also determine whether WM performance reflects age-related increases in controlled attention and storage when predicting performance on fluid and crystallized measures of intelligence. The theoretical model is shown in Figure 3. Except for the direct paths from WM to measures of intelligence, the model was similar to my theoretical model in Figure 2.<sup>6</sup>

As shown at the bottom of Table 3, the theoretical model was not a good fit to the data. The NNFI was below .90. A Lagrange test of this model indicated that a better fit would occur if paths were established between naming speed and fluency and naming speed and random generation. These modifications were added one at a time. When these paths were included in the modeling, there was a good fit of the model to the data. Overall, the revised model made theoretical sense because both the fluency and random generation measures involved time constraints. A comparison between the revised model and the theoretical model was significant (228.28–201.64),  $\Delta\chi^2(2) = 26.59$ ,  $p < .001$ . The standardized parameters for this model are shown in Figure 4. To simplify this diagram, I only present the standardized solutions for the latent

<sup>6</sup> Prior to the analysis, I determined whether it was necessary to eliminate extreme outliers from the path analysis. I calculated the estimated Mardia's Kurtosis [ $p(p + 2)$ , where  $p$  is the number of variables (19 in this case)] as 399. Mardia's estimate of Kurtosis was 14.72 for this sample and therefore I have multivariate data. The Mardia-based Kappa was .039. I also inspected the intercorrelations among the latent and manifest variables and these are shown in Table 4. As shown, the estimated correlation between the latent variables of phonological knowledge and reading are so high ( $r = .90$ ) that they are redundant. Two basic ways to handle this multicollinearity issue is to either eliminate these variables or combine them. Because I were interested in predicting reading, phonological knowledge was deleted from the subsequent analysis.



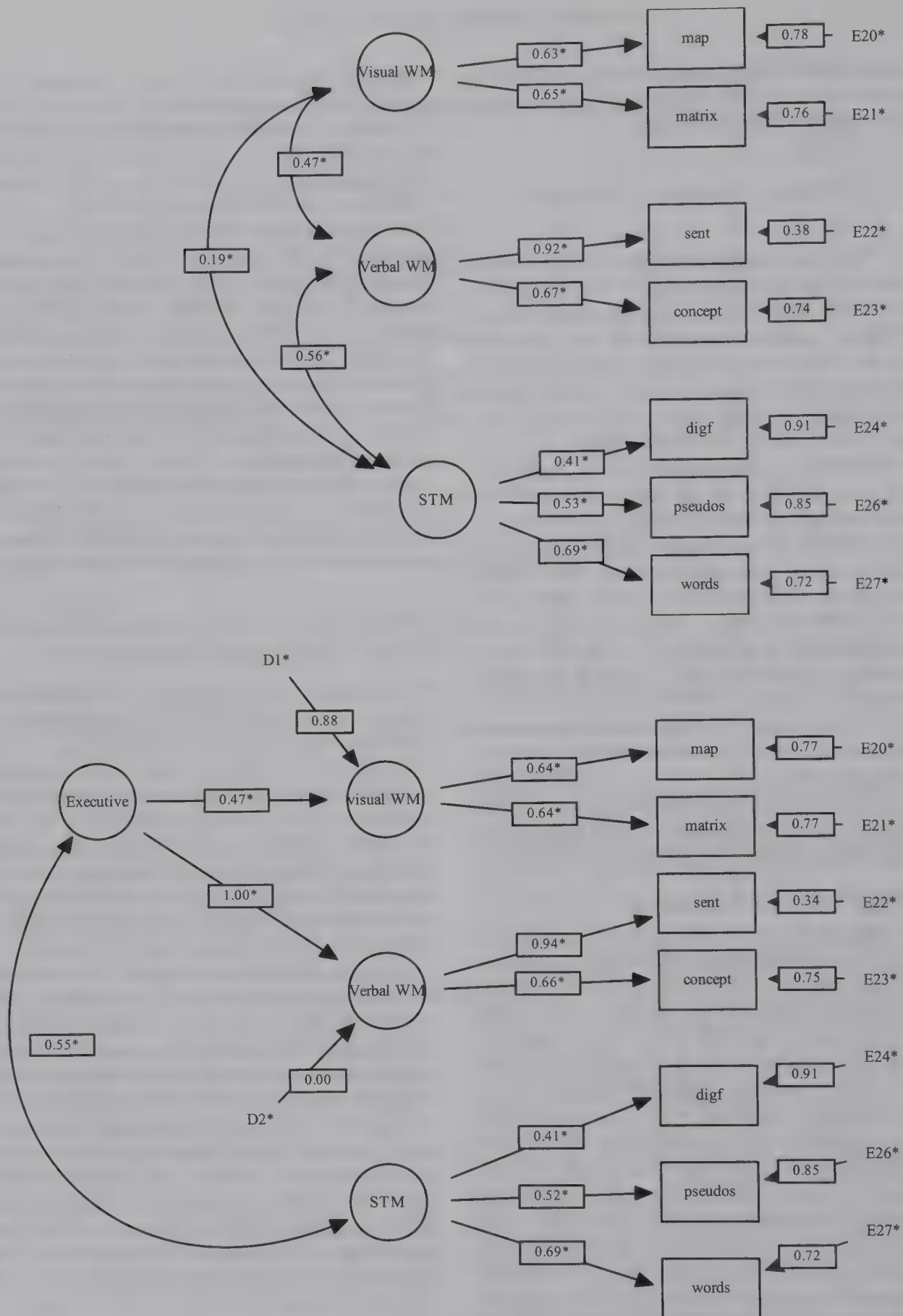


Figure 1. Confirmatory model for a correlated three-factor structure (top) and a second-order factor (bottom). Circles represent latent variables, and squares represent observed or manifest variables. Numbers on paths (path coefficients for the standardized solution) leading from latent variables to observed variables indicate the degree to which an observed variable was influenced by a particular latent variable. All path coefficients with an asterisk were significant at the .05 level. Paths leading to the observed variable that are not attached to latent variables are the residuals (or the degree to which the observed variable was influenced by unique factors). The top panel shows that the standardized factor loadings for the second-order model ranged in size from .41 to .94. The correlation between the latent measure of the second-order factor (referred to as the executive system) and STM storage (also referred to as the phonological loop) was .55. WM = working memory; STM = short-term memory; map = Mapping and Directions; matrix = Visual Matrix; sent = Listening/Sentence Span; concept = Conceptual Span; digif = Forward Digit; pseudos = Pseudoword Span; words = Word Span; D = disturbance; E = measurement error; \* = significance.

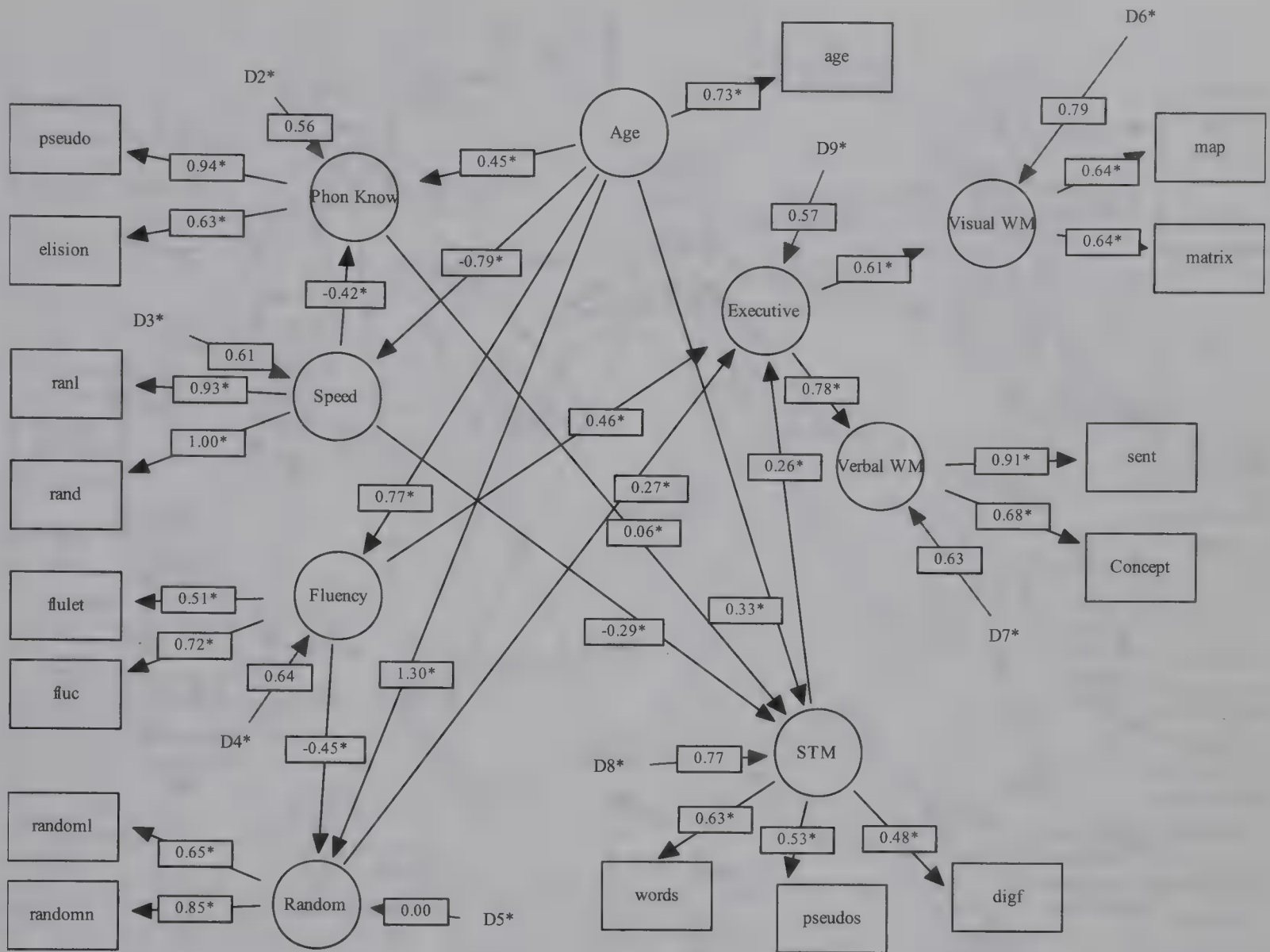


Figure 2. Path analysis to determine age-related correlates of working memory (WM) development. digitf = Forward Digit; pseudos = Pseudoword Span; words = Word Span; map = Mapping and Directions; matrix = Visual Matrix; sent = Listening/Sentence Span; Concept = Conceptual Span; Age = age in months (latent variable); age = age in months (manifest variable); pseudo = pseudoword reading task; elision = Elision or phonological deletion; ranl = rapid naming of letters; rand = rapid naming of digits; flulet = letter fluency; fluc = category fluency; randoml = random generation of letters; randomn = random generation of numbers; Phon Know = Phonological knowledge; Speed = Naming Speed; Random = random generation; STM = short-term memory; Executive = shared variance of verbal and visual WM tasks; D = disturbance; E = measurement error; \* = significance.

variables. Error terms and variables that made up the latent measures are not shown. The  $R^2$  for each direct path for each latent variable and the variables included in the direct paths (shown in the parenthesis) was as follows: fluid intelligence (executive processing + age) was 1.0, math (executive processing + age) was .98, reading (STM, executive process, age) was .94, naming speed (age) was .39, fluency (age) was .41, random generation (age) was .75, executive processing (fluency, STM, random generation) was .62, and STM (naming speed) was .57.

In terms of the magnitude of the standardized path coefficients, Figure 4 shows that the effect size was large for a path from executive processing to fluid intelligence (standardized path coefficient = .77) but moderate in magnitude when predicting math and reading (.30 and .42, respectively). As in the age-related model for WM, the paths from STM, random generation, and fluency to executive processing were significant, but the magnitude of the

coefficients were moderate to small (.10 for STM, .36 for random generation, and .49 for fluency).

Table 4 shows the indirect effects of the modeling. Indirect effects involve intervening variables (mediator variables) assumed to transmit some of the causal effect onto the endogenous variables (Kline, 2005, p. 65). Reported are the nonstandardized and standardized solutions for the decomposition of the paths. There are three important results. First, naming speed played a significant mediating role in fluid and crystallized intelligence. Second, increases in age, fluency, and random generation played mediating roles on measures of fluid intelligence and math. Finally, no significant indirect effects were found for STM.

I next used a hierarchical analysis to determine the mediating role of variables when sequentially entered into the analysis. The correlation matrix for this analysis is shown in Table 5. This was done because I wanted to systematically examine the age-related



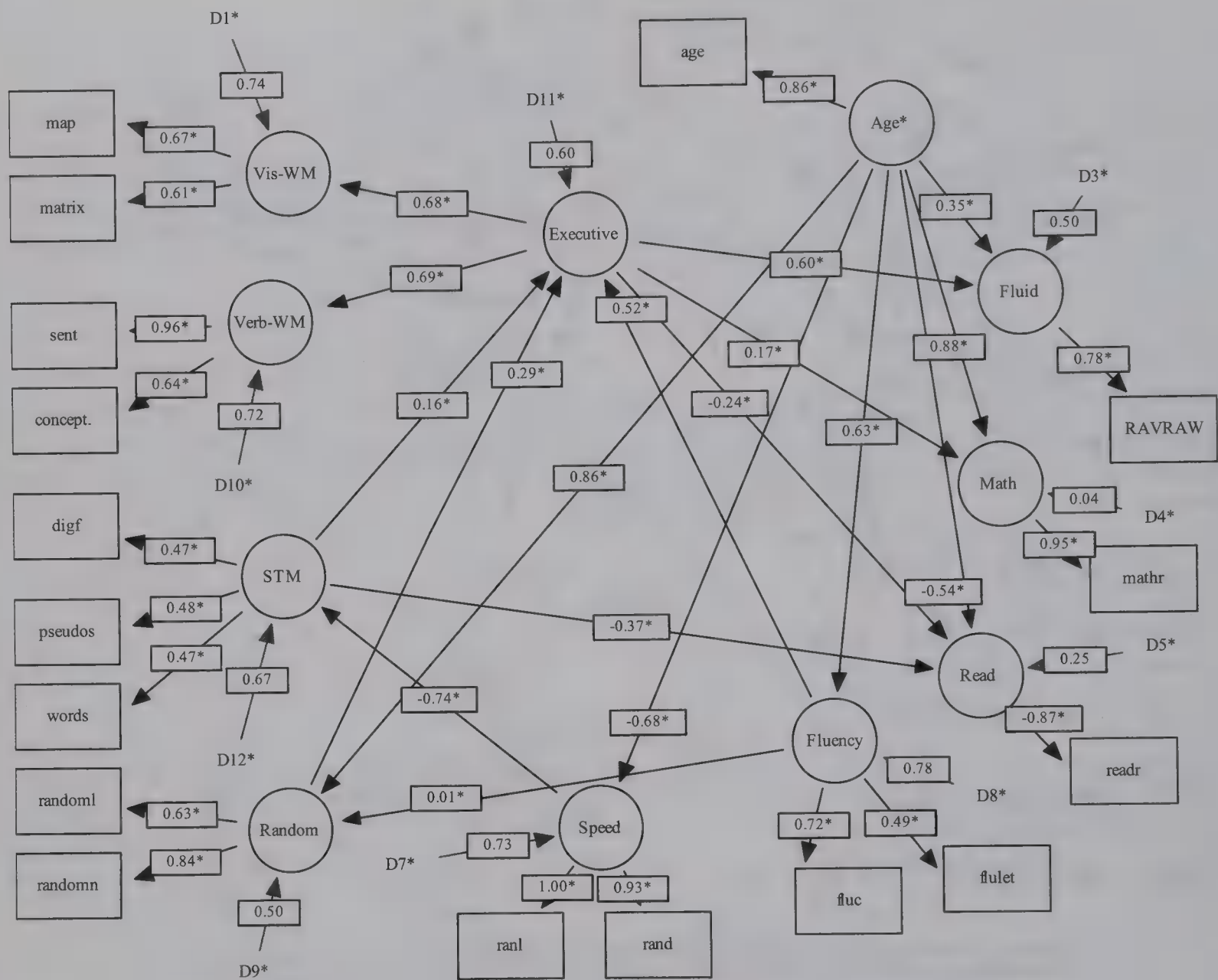


Figure 3. Path analysis representing the theoretical model to determine age-related contributor to fluid and crystallized intelligence. Speed = Naming Speed; Random = random generation; WM = working memory; STM = short-term memory; Executive = shared variance of verbal and visual WM tasks; Fluid = Raven Colored Progressive Matrices; Math = Wide Range Achievement Test—Third Edition Math subtest; Read = Wide Range Achievement Test—Third Edition Reading subtest; map = Mapping and Directions; matrix = Visual Matrix; sent = Listening/Sentence Span; Concept = Conceptual Span; digtf = Forward Digit; pseudos = Pseudoword Span; words = Word Span; randoml = random generation of letters; randomn = random generation of numbers; ranl = rapid naming of letters; rand = rapid naming of digits; flulet = letter fluency; fluc = category fluency; Verb = verbal; Vis = visual; readr = reading raw score; mathr = math raw score; RAVRAW = Raven raw score; D = disturbance; E = measurement error; \* = significance.

variance in fluid and crystallized intelligence, before and after controlling the variance ( $R^2$ ) associated with the introduction of a potential mediating variable (e.g., WM). The difference between the estimates of age-related performance given by the change in  $R^2$  shows the contribution of the mediating variable (e.g., WM) to the age-related differences in fluid and crystallized intelligence (see Salthouse, 2000, for a further rationale). Because this difference (increment in  $R^2$  or  $\Delta R^2$ ) was directly related to the degree to which age-related variance in intelligence was reduced when the mediator was controlled (e.g., WM), I then have a measure of the importance of that mediator. If significant residual age-related

differences remained after the entry of the postulated variable, then the age-related effect “represents influences that are unique to the target variable because the shared influences are presumably partialled out by the control of the other variable” (Salthouse, 2000, p. 4). Criterion measures and predictor variables were latent measures and manifest variables (e.g., age, fluid and crystallized intelligence) from the path analysis.

For my first set of analyses, I determined the amount of variance in fluid intelligence, math, and reading performance that was accounted for by (a) age alone (Model 1) and (b) age after WM and STM was partialled out (Model 2). I next tested (Model 3) whether

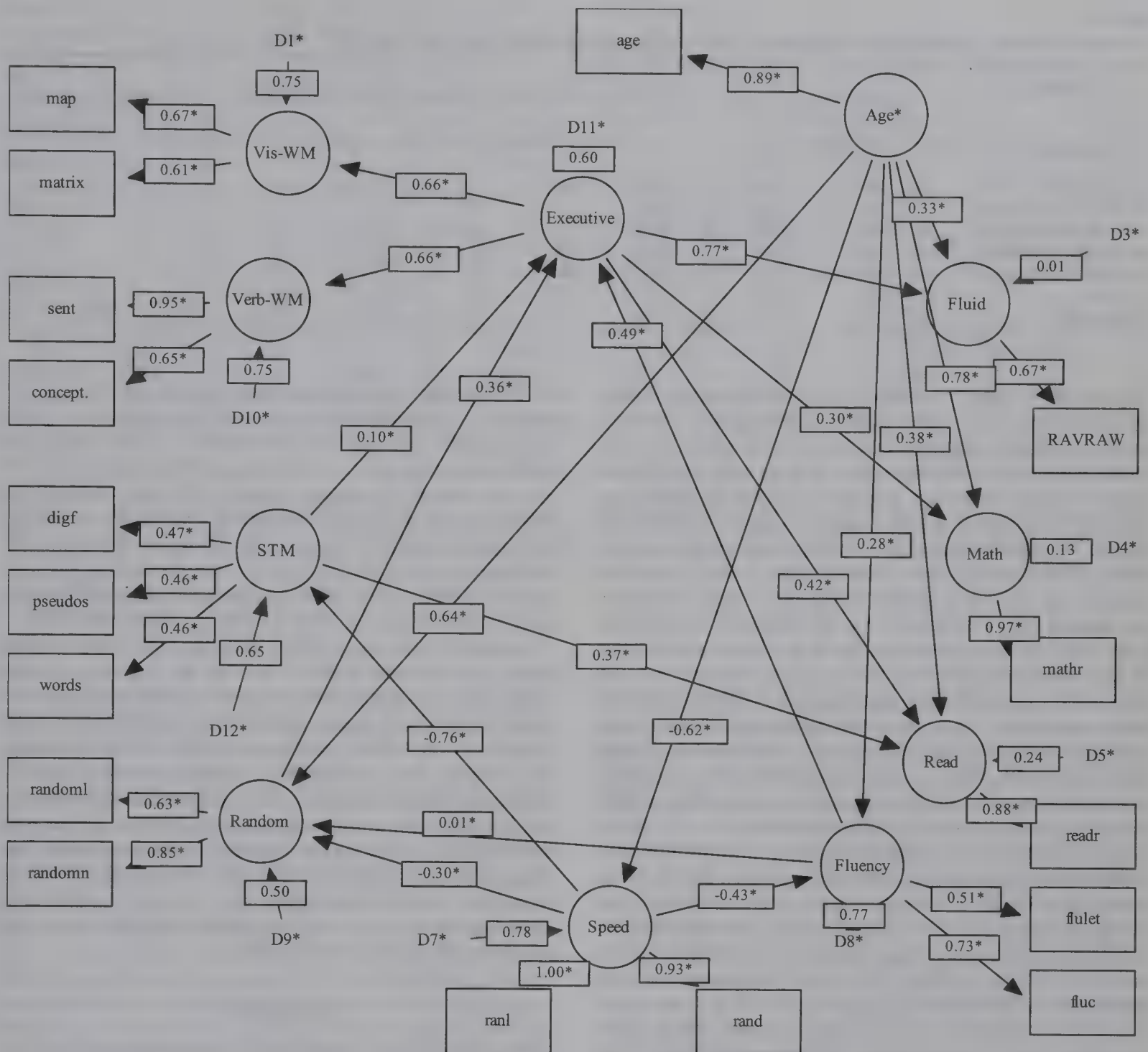


Figure 4. Revised model to determine age-related contributor to fluid and crystallized intelligence. Speed = Naming Speed; Random = random generation; WM = working memory; STM = short-term memory; Executive = shared variance of verbal and visual WM tasks; Fluid = Raven Colored Progressive Matrices; Math = Wide Range Achievement Test—Third Edition Math subtest; Read = Wide Range Achievement Test—Third Edition Reading subtest; map = Mapping and Directions; matrix = Visual Matrix; sent = Listening/Sentence Span; Concept = Conceptual Span; digf = Forward Digit; pseudos = Pseudoword Span; words = Word Span; randoml = random generation of letters; randomn = random generation of numbers; ranl = rapid naming of letters; rand = rapid naming of digits; fulet = letter fluency; fluc = category fluency; Verb = verbal; Vis = visual; readr = reading raw score; mathr = math raw score; RAVRAW = Raven raw score; D = disturbance; E = measurement error; \* = significance.

the significant variance related to WM (with the influence of STM partialled out) on measures of intelligence was eliminated when measures of the phonological system (naming speed) or activities related to executive processing (fluency, inhibition) were entered into the model. The results of my regression analysis for Models 1–3 are presented in Table 6. For each

criterion measure, the cumulative  $R^2$  associated with the addition of variables entered into the regression equation is presented at the bottom of the table. The unique variance in  $R^2$  associated with each variable when entered last in the equation appears in the 5th, 10th, and 15th columns, for fluid intelligence, math, and reading, respectively.



Table 4  
Correlations Among Latent and Manifest (Age, Fluid Intelligence, Math, Reading) Variables ( $N = 205$ )

Variable	1	2	3	4	5	6	7	8	9	10
1. Age	—									
2. Phonological	.60	—								
3. Speed	-.56	-.81	—							
4. Fluency	.55	.79	-.77	—						
5. Random Generation	.79	.80	-.83	.73	—					
6. Short-term memory	.48	.69	-.74	.77	.70	—				
7. Working memory	.49	.61	-.50	.75	.61	.72	—			
8. Fluid Intelligence	.52	.65	-.47	.60	.49	.47	.54	—		
9. Math	.65	.62	-.51	.59	.72	.46	.50	.47	—	
10. Reading	.67	.90	-.81	.78	.83	.68	.59	.64	.68	—

Inspection of Model 1 in Table 6 indicated that age contributed significant variance to all measures of intelligence. The total amount of variance related to the age variable in predicting intelligence was 27% for fluid intelligence, 42% for math, and 45% for word recognition (Model 1). The second analysis (Model 2) that entered WM and STM into the regression model accounted for 38%, 47%, and 62% of the variance in fluid intelligence, math, and reading, respectively. When Models 1 and 2 were considered together, the variance accounted for by age in fluid intelligence was reduced from 27% (Model 1) to 8% (Model 2). More specifically, there was approximately a 70% drop in the influence of age on fluid intelligence following the statistical control of WM and STM [ $.703 = (.27-.08)/.27$ ]. Likewise, entry of the WM and STM latent measure into the regression model reduced the age-related variance in math by 52% [ $.523 = (.42-.20)/.42$ ] and reading by 71% [ $.404 = (.45-.13)/.45$ ]. As shown in Table 6, WM contributed significant variance to fluid intelligence and math, whereas STM contributed unique variance to word recognition.

As shown in Table 6, Model 3 simultaneously entered fluency, inhibition, and naming speed into the regression model. When compared to Model 2, adding these processing variables to the model improved predictions by 6%, 12%, and 16% for fluid intelligence, math, and reading, respectively. The incremental improvement of  $R^2$  from Model 2 to Model 3 was significant when predicting fluid intelligence,  $F_{\text{inc}}(3, 198) = 10.00, p < .001$ ; math,  $F_{\text{inc}}(3, 198) = 20.80, p < .01$ ; and reading,  $F_{\text{inc}}(3, 198) = 53.00, p < .0001$ . Model 3 shows that the significant contribution of WM and age on fluid intelligence remained after entering STM and activities related to controlled attention, whereas the contribution

of STM and WM was eliminated when predicting performance on measures of crystallized intelligence. Significant predictors across all dependent measures were performance on the fluency and random generation measures. Inspection of Table 6 also shows that the total amount of variance related to the age in predicting intelligence was 6% for fluid intelligence, 1% for math, and 2% for word recognition. Thus, equations for Model 3 reduced the contribution of age by 78% ( $.27-.05/.27$ ) when predicting fluid intelligence, by 99% ( $.42-.01/.42$ ) when predicting math performance, and by 96% ( $.45-.02/.45$ ) when predicting reading performance.

In summary, the results yielded three important findings. First, Model 2 that included WM/STM reduced the variance accounted for by age in predicting fluid intelligence, math, and reading. I assumed that reductions greater than 50% were of important note. Reductions in age-related variance were 70% for fluid intelligence, 52% for math, and 71% for reading. Second, as shown in Model 2, the residual variance related to WM predicted fluid intelligence and math, but not word recognition. In contrast, the residual variance related to STM predicted word recognition. Finally, entering of the processing variables into the equation (Model 3) eliminated the significant contribution of memory variables in predicting measures of crystallized intelligence. This was not the case for predicting fluid intelligence.

## Discussion

This study assessed whether STM and WM operated as distinct processes in younger and older children and whether the processes of one system (e.g., phonological) superseded another (executive)

Table 5  
Indirect Effects When Predicting Measures of Intelligence

Causal variable: Indirect effects	Endogeneous variables					
	Fluid Intelligence		Math		Reading	
	Unstandardized	SE	Unstandardized	SE	Unstandardized	SE
Speed	-1.24** (-.29)	.40	-.19** (-.11)	.06	-2.06* (-.44)	1.01
Fluency	1.22** (.37)	.46	.20* (.14)	.07	.74 (.20)	.46
Random	.58** (.27)	.28	.09* (.10)	.04	.35 (.14)	.25
Short-term memory	1.98 (.07)	2.80	.31 (.02)	.44	1.20 (.04)	1.67
Age	1.13** (.40)	.41	.22* (.18)	.06	1.43 (.43)	.73

Note. Values in parentheses are the standardized solution.

\*  $p < .05$ . \*\*  $p < .01$ .

Table 6  
Simultaneous Regression Model for Fluid Intelligence, Math, and Reading ( $N = 205$ )

Model and variable	Fluid Intelligence					Math					Reading				
	<i>B</i>	<i>SE</i>	$\beta$	<i>t</i>	$R^2$	<i>B</i>	<i>SE</i>	$\beta$	<i>t</i>	$R^2$	<i>B</i>	<i>SE</i>	$\beta$	<i>t</i>	$R^2$
Model 1															
Age	.51	.05	.52	8.83***	.27	.75	.06	.65	12.19***	.42	.66	.05	.67	13.03***	.45
Model 2															
STM	.07	.09	.06	.76	.001	.07	.10	.05	.70	.001	.49	.07	.41	6.45***	.07
WM	.35	.08	.33	4.00***	.05	.26	.09	.21	2.77**	.03	.09	.06	.08	1.34	.001
Age	.32	.06	.33	5.16***	.08	.60	.07	.52	8.63***	.20	.42	.04	.42	8.49***	.13
Model 3															
Fluency	.41	.12	.36	3.20**	.03	.47	.13	.34	2.60**	.02	.28	.08	.25	3.54***	.02
Random	-.33	.15	-.32	-2.21***	.02	1.02	.15	.82	6.66***	.09	.30	.09	.29	3.14***	.02
Speed	-.16	.12	-.16	-1.32	.01	.45	.12	.38	3.57***	.03	-.31	.07	-.31	-4.02***	.03
STM	-.10	.11	-.09	-.90	.001	-.20	.12	-.13	-1.67	.01	-.03	.07	-.03	-.48	.001
WM	.26	.10	.19	2.55**	.03	-.01	.10	-.01	-.08	.001	.03	.06	-.03	.55	.01
Age	.40	.09	.41	4.42***	.06	.10	.09	.09	1.14	.01	.12	.05	.12	2.12*	.02

## Model significance

Model 1	$R^2 = .27, F(1, 203) = 78.04, p < .001$	$R^2 = .42, F(1, 203) = 148.67, p < .001$	$R^2 = .45, F(1, 203) = 169.67, p < .001$
Model 2	$R^2 = .38, F(3, 201) = 41.93, p < .001$	$R^2 = .47, F(3, 201) = 59.58, p < .001$	$R^2 = .62, F(3, 201) = 113.03, p < .001$
Model 3	$R^2 = .44, F(6, 198) = 32.81, p < .0001$	$R^2 = .59, F(6, 198) = 48.37, p < .001$	$R^2 = .78, F(6, 198) = 117.58, p < .001$

Note. STM = short-term memory; WM = working memory.

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

in mediating the influence of age-related changes in WM on measures of intelligence. I briefly review the models and the questions that directed the study.

Two models were tested. One model tested whether processes related to the phonological system (phonological loop) played a major role in predicting age-related performance on WM measures and whether the phonological system mediated the influence of WM on measures of intelligence. The model follows logically from literature on STM that links phonological skills to various aspects of children's crystallized intelligence (e.g., new word learning; Baddeley et al., 1998), comprehension (Perfetti, 1985), and arithmetic (e.g., Geary et al., 1999). This model suggests that because younger children are less efficient in the processing of phonological information, this may blur dissociations between measures of the phonological loop and the executive system. Thus, both WM and STM measures load on the same factor in younger children. The model assumes that phonological processing is more likely to mediate age-related differences in children's WM, fluid, and crystallized intelligence than measures of executive processing. I broadly defined phonological processing in this study as performance on latent measures of STM, phonological knowledge, and rapid naming speed. I broadly defined executive processing as performance on latent measures of WM, fluency, and random generation.

The second model posits that the influence of WM performance on measures of fluid and crystallized intelligence is due to executive processing, independent of the influence of the phonological system. The model predicts that regardless of age, WM and STM tasks load on separate factors because they draw on distinct systems: controlled attention and storage. Thus, I expected that STM and WM would operate as independent systems in terms of their influence on measures of fluid and crystallized intelligence. I assumed that controlled attentional processing was tapped when

measures of STM were partialled from the WM scores (e.g., Engle et al., 1999). I expected that WM would better predict measures of intelligence than STM. This assumption follows logically from the literature on fluid intelligence, reading, and math suggesting that abstract thinking requires the coordination of several basic processes (e.g., Engle et al., 1999; Just, Carpenter, & Keller, 1996; Kyllonen & Christal, 1990). I assumed that controlled attentional processing was related to individual differences in inhibition (fluency and random generation). Thus, I tapped two aspects of controlled attentional processing: one focused on the search for words while inhibiting repetitions and the other focused on inhibiting well-learned sequences.

In general, the results give support to the second model. I found that controlled attention moderated age-related changes in WM as well as the influence of WM on measures of intelligence. I found support for the influence of controlled attention (inhibition), as well as STM, in predicting age-related changes in WM. For the present sample, measures of inhibition and storage (STM latent score) contribute unique variance to WM. Thus, I find that the WM variable was related to measures of inhibition independent of STM storage. I now address three questions that directed this study.

### 1. Are WM and STM Distinct Processes in Young Children?

Our results show that STM and WM performance in elementary school children is distinguishable. The structure between the two types of tasks differs. The results show that WM and STM tasks load on separate factors in both younger and older children. The results further show that although STM (phonological loop) and WM (executive system) are correlated ( $r = .55$ ); they can and do operate independently of one another in predicting performance on measures of fluid and crystallized intelligence. More important,



data within and across the two age groups show comparable factor structures. A multisample analysis of the covariance structure showed a good fit suggesting that WM and STM measures are tapping comparable constructs across age groups.

The results of this study also show that the factor structure among children in the elementary grades was stable as well as comparable to the factor structure for STM and WM tasks found in adults (e.g., Engle et al., 1999). In terms of stability, I did not find a qualitative change in the structure of the memory tasks between the two age groups. A three-factor model captured performance in both the younger and older groups of children. I tested whether a one-factor model might have been more adequate for the older than the younger age, but I did not find support for either of these models. When generalizing to adult samples, as found in the Engle et al. (1999) study, I also found that letter, digit, and word span tasks define the STM factor, whereas tasks that include the combination of both processing and storage define the executive component of the WM factor. I defined the processing component of my WM measures as that part of the memory task that directed children to respond to discrimination questions about recently encoded information, whereas I defined the storage component as that part of the task that required children to retrieve recently encoded information. In sum, my findings do not support studies showing WM and STM as loading on a similar factor in children (e.g., Hutton & Towse, 2001).

The above findings are important because Engle et al. (1999) suggested that rehearsal and chunking may play a role in the poor differentiation of STM and WM tasks in children. They suggested that because these processes are less automated in children, then perhaps an executive system might be invoked on STM tasks that in turn would make STM and WM less distinct. I did not find direct support for this hypothesis. In fact, the correlations between STM and WM were substantially weaker than that of the Engle et al. study ( $r = .70$  vs.  $.55$ ) suggesting greater independence among these measures in children than adults.

## 2. Do Age-Related Improvements in Controlled Attention and/or Storage Underlie Age-Related Changes in Children's WM?

The path analysis found that approximately 70% of the variance in WM was attributable to STM storage, fluency, and random generation. Thus, the results clearly support the notion that the development of WM is related to both increases in storage and controlled attention in young children. The path analysis, however, gave us some unique insights in how these age-related changes influence WM performance. Naming speed had a significant path to STM, suggesting that naming speed was important to storage. Thus, although age-related changes in controlled attention activities (random generation, fluency) and STM have a direct path to WM, STM storage is influenced by age-related increases in phonological knowledge and naming speed. Although phonological knowledge had a significant path to STM, the effect size for this path was substantially smaller than naming speed.

Overall, these results on WM are consistent with recent studies that have shown that age-related changes in WM are related to storage variables (Bayliss et al., 2005). However, WM is made up of more than just storage. More specifically, Bayliss et al. (2005) found that "age-related improvements in storage ability were important for complex span independent of

any changes in processing speed and conversely, that age-related changes in speed of processing contributed to complex span performance independent of changes in storage ability" (p. 592). I found clear support that age-related changes in STM predicted age-related changes in WM. Although my measure of speed (naming speed) differed from that of Bayliss et al. (2005; reaction time of verbal and visual-spatial search tasks), and therefore a direct comparison between the two studies cannot be made, my results suggest that speed had a direct effect on STM storage and not on WM. Further, I were able to specify that storage is made up of two components: phonological knowledge and naming speed. More important, in contrast to the other studies in which speed predicts WM, I did not find that naming speed played a mediating role on WM performance. However, as shown in Table 4, latent measures of naming speed and STM/WM were moderate ( $-.74$  and  $-.50$ , respectively), and in the same range as raw scores in the Bayliss et al. (2005) study (see Table 2; correlations between storage tasks and basic speed measures were  $-.46$  and  $-.47$ ). However, my analysis had the advantage of including measures of controlled attention. In general, my results are in line with studies using adult samples that have found verbal storage in STM accounts for age-related effects on WM measures. For example, McCabe and Hartman's (2003) results showed that storage capacity (STM) in language processing was an important predictor of complex WM performance. However, as an extension of these studies I found that processes related to controlled attention (inhibition), along with STM, played the major role in predicting age-related changes in WM. STM, fluency, and random generation each contributed unique variance to WM. The path analysis showed that approximately 70% of the WM variance could be attributed to STM and controlled attention measures.

In summary, the results show that WM in children is made up of controlled attention and STM, whereas STM is made up of naming speed and phonological knowledge.

## 3. What Processes Influence WM on Age-Related Changes in Intelligence?

Answering this question was important because the processes that mediate age-related development in WM span may not be the same processes that mediate the influence of WM on intelligence. In my model testing, I determined whether the theoretical model that fit the processes that mediate WM span also play an important role when WM was used to predict fluid and crystallized intelligence. That is, the theoretical model assumed that age-related changes in controlled attention activities and storage had a direct influence on WM span that in turn had a direct effect on intelligence. Unfortunately, the aforementioned model was not a good fit to the data when predicting fluid and crystallized intelligence and was revised. The revised model found that direct effects from naming speed to inhibition measures were necessary. Thus, naming speed played a significant mediating role between WM and measures of intelligence.

When the analysis focused on fluid and crystallized intelligence in isolation, the hierarchical regression analysis showed that the residual variance related to WM predicted fluid intelligence and math, whereas the residual variance related to STM predicted reading. More important, there was approximately a 70% drop in the influence of age on fluid intelligence, a 52% drop in math, and



a 71% drop in reading when measures of STM and WM were entered into the analysis (see Model 2). The results in Table 6 also showed that for the full model, entry of controlled attention measures into the regression model eliminated the significant contribution of WM to measures of crystallized intelligence, but not fluid intelligence. Thus, the question emerges as to whether the residual variance attributed to WM and fluid intelligence reflects controlled attention (e.g., Engle et al., 1999). I assumed that controlled attention was related to the latent measures of fluency and random generation. There is good reason to assume this because the correlation between the latent measures of WM and fluency was .75 and between WM and random generation was .61 (see Table 4). Further, there is good reason to assume from the literature that the measures tap some aspects of controlled attention, such as the inhibition of well-learned sequences (e.g., Towse, 1998). However, I found that entry of these measures in the regression models eliminated the significant contribution of WM only on measures of math and reading, but not on fluid intelligence. What these findings suggest to us is that the residual variance related to WM (controlled attention) in predicting fluid intelligence is mediated by processes other than inhibition. Further, because my measures were not pure measures of inhibition, I may not have adequately tested the inhibition model. Taken together, however, my results extend Engle et al.'s (1999) findings to children that when controlling for the correlations between WM and STM, the residual variance for the WM factor predicts fluid intelligence.

### Implications

There are two implications related to my findings. First, age-related changes in WM are made up of more than storage and speed. Measures of controlled attention, in this case inhibition, must also be included in the modeling. Some developmental models have indicated that younger children are less resistant to interference than older children (see Bull & Scerif, 2001; Brainerd & Reyna, 1993; Chiappe, Hasher, & Siegel, 2000; De Beni, Palladino, Pazzaglia, & Cornoldi, 1998). There are also several memory studies with adult samples showing that inhibitory capabilities are a determinant of WM differences. For example, Hasher, Stoltz, Zacks, and Ryma (1991) have shown that age differences in WM and comprehension in language are driven by an inhibitory deficit. Older adults show difficulties on WM tasks not because of a smaller capacity but because an inhibitory mechanism fails to regulate the contents of WM. Failure comes when relevant and irrelevant information competes for retrieval and control. In Hasher, Lustig, and Zacks's (2007) view, then, inhibition drives WM capacity. When applying this model to age-related changes in children's WM, some researchers argue that poor inhibition underlies younger children's ability to prevent irrelevant information from entering WM during the processing of targeted information (see Passolunghi & Siegel, 2001, for a discussion of this model). Younger children have difficulty preventing unnecessary information from entering WM and, therefore, are more likely to consider alternative interpretations of material than older children. This interpretation fits with my findings, as well as several models with adults that explain individual differences in WM performance as related to inhibitory mechanisms (e.g., Cantor & Engle, 1993; Conway & Engle, 1994; also see Barrett, Tugade, & Engle, 2004, for a review).

In support of this model, I found that inhibition (fluency and random generation) measures predict WM performance as well as performance on measures of fluid and crystallized intelligence. Thus, WM tasks tap the child's ability to activate relevant information as well as suppress irrelevant information that in turn places demands on a general capacity system. Although I see the inhibition model as one of several processes underlying the residual variance related to WM, I do not see it as the sole process underlying age-related differences in WM. A special account has to be given to the role of other processes, especially because WM predicted fluid intelligence when all measures of inhibition were entered into the regression model.

The findings also give support to a popular model of WM related to speed of processing. A simple version of this model states that young children are slower at processing language information than older children and that such reduced processing on the younger participants' part underlies their poor WM performance. Several models of WM assume that operations related to language are time consuming (e.g., Ackerman, Beier, Boyle, 2002; Salthouse, 1996; Verhaeghen & Salthouse, 1997). Therefore, speed of processing may underlie the general pattern of age-related changes in WM noted in the present study. Further, Kail (1993) argued that a common pool of cognitive resources that relate to processing speed are used to perform a variety of tasks, with the pool increasing across ages in children. Clearly, my findings show a significant correlation between latent measures of naming speed and latent measures of STM and WM ( $r = -.74$  and  $-.50$ , respectively; see Table 4). However, I found that naming speed is primarily related to the storage component of WM. It is important to note, however, that my study focused on naming speed and therefore the results may have been different if measures of speed related to decision making, reaction time, and related measures would have been measured.

Second, the relationship between the basic cognitive capacities of STM, WM, and processing speed with measures of fluid and crystallized intelligence was substantial. Latent measures of basic capacities were each found to correlate substantially with the intelligence measures, and the basic cognitive factors in the SEM accounted for over 93% of the variance in fluid and crystallized intelligence. The results suggest a strong link between children's intelligence and basic cognitive processes. The study replicates previous studies on the importance of basic cognitive processes in predicting academic performance. For example, Luo, Thompson, and Detterman (2006) in their recent large-scale analyses found similar relations between the basic cognitive factors of STM, WM, and processing speed and a scholastic performance factor. The variability in scholastic performance explained by the basic capacities was found to be close to 90% or more on the level of latent factors.

### Conclusion

In conclusion, the present results can address the question "What does the development of WM in young children entail?" The development of WM involves two major components: controlled attention and storage. I found that age-related improvements in WM are significantly related to individual differences in controlled attention and STM, whereas phonological knowledge and naming speed mediate STM storage. The two systems, while related, can operate independent of one another. Our findings



indicate that components of STM and WM load on distinct factors in children as young as 6 years of age. I also found that when WM is used in predicting measures of intelligence, naming speed plays a significant mediating role. That is, while WM growth is related to controlled attention and storage, other processes, such as naming speed, come into play when mediating the role of WM on measures of intelligence.

## References

- Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2002). Individual differences in working memory within a nomological network of cognitive and perceptual speed abilities. *Journal of Experimental Psychology: General*, 131, 567–589.
- Alloway, T. P., Gathercole, S. E., Willis, C., & Adams, A. (2004). A structural analysis of working memory and related cognitive skills in young children. *Journal of Experimental Child Psychology*, 87, 85–106.
- Baddeley, A. D. (1986). *Working memory*. London: Oxford University Press.
- Baddeley, A. D. (1996). Exploring the central executive. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 49(A), 5–28.
- Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, 4, 417–423.
- Baddeley, A. D., Gathercole, S. E., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, 105, 158–173.
- Baddeley, A. D., Lewis, V. J., & Vallar, G. (1984). Exploring the articulatory loop. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 36(A), 223–252.
- Baddeley, A. D., & Logie, R. H. (1999). Working memory: The multiple-component model. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 28–61). New York: Cambridge University Press.
- Barrett, L. F., Tugade, M. N., & Engle, R. (2004). Individual differences in working memory capacity and dual-process theories of the mind. *Psychological Bulletin*, 130, 553–573.
- Barrouillet, P., & Camos, V. (2001). Developmental increase in working memory span: Resource sharing or temporal decay? *Journal of Memory and Language*, 45, 1–20.
- Bayliss, D. M., Jarrold, C., Baddeley, A. D., Gunn, D. M., & Leigh, E. (2005). Mapping the developmental constraints on working memory span. *Developmental Psychology*, 41, 579–597.
- Bayliss, D. M., Jarrold, C., Gunn, D. M., & Baddeley, A. (2003). The complexities of complex span: Explaining individual differences in working memory in children and adults. *Journal of Experimental Psychology: General*, 132, 71–92.
- Bentler, P. M. (2006). *Structural equations program manual (EQS 6.1)*. Encino, CA: Multivariate Software.
- Bentler, P. M., & Bonett, D. (1980). Significance test and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588–606.
- Bentler, P. M., & Wu, E. J. C. (1995). *Structural equations program manual*. Encino, CA: Multivariate Software.
- Brainerd, C. J., & Reyna, V. F. (1993). Domains of fuzzy trace theory. In M. L. Howe & R. Pasnak (Eds.), *Emerging themes in cognitive development: Vol. 1. Foundations* (pp. 50–93). New York: Springer-Verlag.
- Bull, R., & Scerif, G. (2001). Executive functioning as a predictor of children's mathematics ability: Inhibition, switching, and working memory. *Developmental Neuropsychology*, 19, 273–293.
- Cantor, J., & Engle, R. W. (1993). Working-memory capacity is long-term memory activation: An individual-differences approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 1101–1114.
- Cantor, J., Engle, R. W., & Hamilton, G. (1991). Short-term memory, working memory, and verbal abilities: How do they relate? *Intelligence*, 15, 229–246.
- Case, R. (1992). *The mind's staircase: Exploring the conceptual underpinnings of children's thought and knowledge*. Hillsdale, NJ: Erlbaum.
- Cattell, R. B. (1971). *Abilities: Their structure, growth and action*. Boston: Houghton Mifflin.
- Cattell, R. B. (1973). *Measuring intelligence with the Culture Fair tests*. Champaign, IL: Institute for Personality and Ability Testing.
- Chiappe, P., Hasher, L., & Siegel, L. S. (2000). Working memory, inhibitory control, and reading disability. *Memory & Cognition*, 28, 8–17.
- Cohen, J. (1988). *Statistical power for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Conlin, J. A., Gathercole, S. E., & Adams, J. W. (2005). Children's working memory: Investigating performance limitations in complex span. *Journal of Experimental Child Psychology*, 90, 303–317.
- Conway, A. R. A., Cowan, N., Bunting, M. F., Theriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, 30, 163–183.
- Conway, A. R., & Engle, R. W. (1994). Working memory and retrieval: A resource-dependent inhibition model. *Journal of Experimental Psychology: General*, 123, 354–373.
- Cowan, N. (1995). *Attention and memory: An integrated framework* (Oxford Psychology Series No. 26). New York: Oxford University Press.
- Cowan, N., Towse, J. N., Hamilton, Z., Saults, J. S., Elliott, E. M., Lacey, J. F., Moreno, M. V., & Hitch, G. J. (2003). Children's working memory processes: A response timing analysis. *Journal of Experimental Psychology: General*, 132, 113–132.
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19, 450–466.
- De Beni, R., Palladino, P., Pazzaglia, F., & Cornoldi, C. (1998). Increases in intrusion errors and working memory deficit of poor comprehenders. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 51(A), 305–320.
- Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General*, 128, 309–331.
- Fry, A. F., & Hale, S. (2000). Relationships among processing speed, working memory, and fluid intelligence in children. *Biological Psychology*, 54, 1–34.
- Gathercole, S. E. (1998). The development of memory. *Journal of Child Psychology and Psychiatry*, 39, 3–27.
- Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language*. Hove, England: Erlbaum.
- Gathercole, S. E., Lamont, E., & Alloway, T. P. (2006). Working memory in the classroom. In S. Pickering (Ed.), *Working memory in the classroom* (pp. 220–238). Oxford, England: Academic Press.
- Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, 40, 177–190.
- Gavens, N., & Barrouillet, P. (2004). Delays of retention, processing efficiency, and attentional resources in working memory. *Journal of Memory and Language*, 51, 644–657.
- Geary, D. C., Hoard, M. K., & Hamson, C. O. (1999). Numerical and arithmetical cognition: Patterns of functions and deficits in children at risk for mathematical disability. *Journal of Experimental Child Psychology*, 74, 213–239.
- Griffiths, Y. M., & Snowling, M. J. (2002). Predictors of exception word and nonword reading in dyslexic children: The severity hypothesis. *Journal of Educational Psychology*, 94, 34–43.
- Haavisto, M., & Lehto, J. E. (2005). Fluid/spatial and crystallized intelligence in relation to domain-specific working memory: A latent-variable approach. *Learning and Individual Differences*, 15, 1–21.

- Hambrick, D. Z., & Engle, R. (2002). Effects of domain knowledge, working memory capacity, and age on cognitive performance: An investigation of the knowledge-is-power hypothesis. *Cognitive Psychology*, 44, 339–387.
- Harrison, J. E., Buxton, P., Husain, M., & Wise, R. (2000). Short-test of semantic and phonological fluency: Normal performance, validity, and test-retest reliability. *British Journal of Clinical Psychology*, 39, 181–191.
- Hasher, L., Lustig, C., & Zacks, R. (2007). Inhibitory mechanisms and the control of attention. In A. Conway, C. Jarrold, M. Kane, A. Miyake, & J. N. Towse (Eds.), *Variations in working memory* (pp. 227–249). New York: Oxford University Press.
- Hasher, L., Stoltz, E. R., Zacks, R. T., & Ryma, B. (1991). Age and inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17, 163–169.
- Heitz, R. P., Unsworth, N., & Engle, R. W. (2005). Working memory capacity, attention control, and fluid intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 61–78). Thousand Oaks, CA: Sage.
- Henry, L. A., & Millar, S. (1993). Why does memory span improve with age? A review of the evidence for two current hypotheses. *European Journal of Cognitive Psychology*, 5, 241–287.
- Hitch, G. J., Towse, J. N., & Hutton, U. (2001). What limits children's working memory span? Theoretical accounts and applications for scholastic development. *Journal of Experimental Psychology: General*, 130, 184–198.
- Horn, J. L. (1980). Concepts of intellect in relation to learning and adult development. *Intelligence*, 4, 285–317.
- Horn, J. L., & Noll, J. (1997). Human cognitive abilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft, & Harrison, P. (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues*. (pp. 53–91). New York: Guilford Press.
- Hulme, C., Thomson, N., Muir, C., & Lawrence, A. (1984). Speech rate and the development of short-term memory span. *Journal of Experimental Child Psychology*, 47, 72–87.
- Hutton, U. M. Z., & Towse, J. N. (2001). Short-term memory and working memory as indices of children's cognitive skills. *Memory*, 9, 383–394.
- Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI: Analysis of linear structural relationships by the method of maximum likelihood*. Mooresville, IN: Scientific Software.
- Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99, 122–149.
- Just, M. A., Carpenter, P. A., & Keller, T. A. (1996). The capacity theory of comprehension: New frontiers of evidence and arguments. *Psychological Review*, 103, 773–780.
- Kail, R. (1993). The role of global mechanisms in developmental change in speed of processing. In M. Howe & R. Pasnak (Eds.), *Emerging themes in cognitive development* (Vol. 1, pp. 97–116). New York: Springer-Verlag.
- Kane, M. J. (2005). Full frontal fluidity? Looking in on the neuroimaging of reasoning and intelligence. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 141–164). Thousand Oaks, CA: Sage.
- Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford Press.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.
- Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity! *Intelligence*, 14, 389–433.
- Luo, D., Thompson, L. A., & Detterman, D. K. (2006). The criterion validity of tasks of basic processes. *Intelligence*, 34, 2006.
- McCabe, J., & Hartman, M. (2003). Examining the locus of age effects on complex span tasks. *Psychology and Aging*, 18, 562–572.
- McDougall, S., Hulme, C., Ellis, A., & Monk, A. (1994). Learning to read: The role of short-term memory and phonological skills. *Journal of Experimental Child Psychology*, 58, 112–133.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525–543.
- Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., & Howerter, A. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, 41, 49–100.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Oberauer, K., Suß, H.-M., Wilhelm, O., & Wittman, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence*, 31, 167–193.
- Passolunghi, M. C., & Siegel, L. S. (2001). Short-term memory, working memory, and inhibitory control in children with difficulties in arithmetic problem solving. *Journal of Experimental Child Psychology*, 80, 44–57.
- Paulesu, E., Demonet, J., Fazio, F., McCrory, E., Chanoine, V., Brunswick, N., et al. (2001, March 16). Dyslexia: Cultural diversity and biological unity. *Science*, 291, 2165–2167.
- Perfetti, C. (1985). *Reading ability*. New York: Oxford Press.
- Psychological Corporation. (1991). *Wechsler Intelligence Scale for Children—Third Edition*. San Antonio, TX: Harcourt Brace Jovanovich.
- Raven, J. C. (1976). *Colored progressive matrices*. London: H. K. Lewis.
- Rende, B., Ramsberger, G., & Miyake, A. (2002). Commonalities and differences in working memory components underlying letter and category fluency tasks: A dual-task investigation. *Neuropsychology*, 16, 309–321.
- Rosen, V. M., & Engle, R. W. (1997). The role of working memory capacity in retrieval. *Journal of Experimental Psychology: General*, 126, 211–227.
- Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review*, 103, 403–428.
- Salthouse, T. A. (2000). Item analysis of age relations on reasoning tests. *Psychology and Aging*, 15, 3–8.
- Schweizer, K., & Koch, W. (2002). A revision of Cattell's investment theory: Cognitive properties influence learning. *Learning and Individual Differences*, 13, 57–82.
- Shankweiler, D., & Crain, S. (1986). Language mechanisms and reading disorder. *Cognition*, 24, 139–168.
- Siegel, L. S. (1994). Working memory and reading: A life-span perspective. *International Journal of Behavioral Development*, 17, 109–124.
- Swanson, H. L. (1992). Generality and modifiability of working memory among skilled and less skilled readers. *Journal of Educational Psychology*, 84, 473–488.
- Swanson, H. L. (1995). *S-Cognitive Processing Test*. Austin, TX: Pro-Ed.
- Swanson, H. L. (1996). Individual and age-related differences in children's working memory. *Memory & Cognition*, 24, 70–82.
- Swanson, H. L. (1999). What develops in working memory? A life span perspective. *Developmental Psychology*, 35, 986–1000.
- Swanson, H. L., & Ashbaker, M. (2000). Working memory, short-term memory, articulation speed, word recognition, and reading comprehension in learning disabled readers: Executive and/or articulatory system? *Intelligence*, 28, 1–30.
- Swanson, H. L., Ashbaker, M., & Lee, C. (1996). Learning disabled readers' working memory as a function of processing demands. *Journal of Experimental Child Psychology*, 61, 242–275.
- Swanson, H. L., & Berninger, V. (1995). The role of working memory and STM in skilled and less skilled readers' word recognition and comprehension. *Intelligence*, 21, 83–108.
- Swanson, H. L., & Siegel, L. (2001). Learning disabilities as a working memory deficit. *Issues in Education: Contributions of Educational Psychology*, 7, 1–43.
- Towse, J. (1998). On random generation and the central executive of working memory. *British Journal of Psychology*, 89, 77–101.



- Towse, J. N., Hitch, G., & Hutton, U. (1998). A reevaluation of working memory capacity in children. *Journal of Memory and Language*, 39, 195-217.
- Verhaeghen, P., & Salthouse, T. A. (1997). Meta-analysis of age-cognition relations in adulthood: Estimates of linear and nonlinear age effects and structural models. *Psychological Bulletin*, 122, 231-249.
- Wagner, R., & Torgesen, J. (1999). *Test of Word Reading Efficiency*. Austin, TX: Pro-Ed.
- Wagner, R., Torgesen, J., & Rashotte, C. (2000). *Comprehensive Test of Phonological Processes*. Austin, TX: Pro-Ed.
- Wagner, T. D., & Smith, E. E. (2003). Neuroimaging study of working memory: A meta-analysis. *Cognitive, Affective, and Behavioral Neuroscience*, 3, 255-274.
- Wilkinson, G. S. (1993). *The Wide Range Achievement Test—Third Edition*. Wilmington, DE: Wide Range.
- Woodcock, R. W. (1998). *Woodcock Reading Mastery Test—Revised (Form G)*. Circle Pines, MN: American Guidance Service.

Received May 22, 2007

Revision received November 15, 2007

Accepted November 19, 2007 ■

### E-Mail Notification of Your Latest Issue Online!

Would you like to know when the next issue of your favorite APA journal will be available online? This service is now available to you. Sign up at <http://notify.apa.org/> and you will be notified by e-mail when issues of interest to you become available!

# When Less Is More in Cognitive Diagnosis: A Rapid Online Method for Diagnosing Learner Task-Specific Expertise

Slava Kalyuga  
University of New South Wales

Rapid cognitive diagnosis allows measuring current levels of learner domain-specific knowledge in online learning environments. Such measures are required for individualizing instructional support in real time, as students progress through a learning session. This article describes 2 experiments designed to validate a rapid online diagnostic method that was inspired by experimental procedures applied in classical cognitive studies of chess expertise. With the described rapid verification method, learners are required to rapidly verify suggested steps at various stages of a problem solution procedure. In this study involving 33 university students, a high degree of correlation was found between rapid testing scores and results of in-depth cognitive diagnosis based on observations of problem-solving steps using video recordings and concurrent verbal reports in the domains of kinematics (vector addition motion problems) and mathematics (transforming graphs of linear and quadratic functions). The article discusses possible applications of the suggested method in adaptive learning environments.

**Keywords:** online cognitive diagnosis, rapid diagnostic method, expertise, learner-tailored instruction

*Task-specific expertise* is the ability of a person to perform fluently in a specific class of tasks. It has a more narrow definition than *professional expertise* (e.g., Ericsson & Charness, 1994), and it is an essential part of academic domain-specific expertise within the model of domain learning (Alexander, 2004). For example, a secondary school student could reach a top level of task-specific expertise in solving simple linear algebra equations, although he or she still would be far from becoming not only an expert mathematician but also an expert in school-level mathematics. Although both professional expertise and academic domain learning expertise include additional essential attributes (e.g., a systemic vision of the field, well-developed metacognitive skills, strategic processes, attitudes and interests), a common major feature of all types of expertise is the availability of a well-organized domain-specific knowledge base (Bransford, Brown, & Cocking, 1999). In the case of task-specific expertise, it is knowledge structures (including strategies and procedures) used in solving a specific class of tasks. Because of the full or partial automation of such structures, experts are able to rapidly perform advanced solution stages by integrating procedures and skipping some (or all) intermediate steps (Blessing & Anderson, 1996; Koedinger & Anderson, 1990; Sweller, Mawer, & Ward, 1983).

Task-specific expertise is a necessary prerequisite of both high-level professional expertise and academic domain learning expertise. Experts of both these types are also highly proficient in solving key classes of tasks within their areas of expertise. Therefore, in education and training, developing expertise in main classes of tasks is an important condition of mastering specific subject domains. Within a cognitive load framework (see Van

Merriënboer & Sweller, 2005, for a recent overview), the importance of developing task-specific expertise is explained by the need to free cognitive resources that are required for learning higher level tasks and developing flexible and transferable skills.

A number of recent studies have demonstrated significant interactions between levels of learner task-specific expertise and optimal instructional formats (expertise reversal effect; for overviews, see Kalyuga, 2005; 2006b; Kalyuga, Ayres, Chandler, & Sweller, 2003). Many instructional procedures and formats that are effective for novice learners may become ineffective, or even detrimental, for expert learners and vice versa. The major instructional implication of the effect is the need to adapt the design of learning environments and levels of instructional guidance to changing levels of learner expertise in a corresponding class of tasks. To be able to do this dynamically, in real time (e.g., during a single tutoring session), appropriate rapid methods of cognitive diagnosis are required that are capable of accurately detecting different levels of learner task-specific expertise. Such rapid methods are especially important for adaptive online learning environments and Web-based courses that cannot use traditional paper-based diagnostic instruments (or their online equivalents) in real time. E-learning courses and tutorials are becoming widespread; however, their adaptive capabilities are usually limited to tailoring instructional content to relatively superficial learner attributes (e.g., preferences, interests, choices, history of previous online behavior) and are not based on fundamental cognitive characteristics such as learner knowledge base.

Traditional methods that are best suitable for diagnosing individual levels of expertise in a domain are based on interviews and think-aloud protocols as learners perform domain-specific tasks. Such verbal reports may provide near-direct evidence about real cognitive processes, ways of reasoning, and underlying learner knowledge structures. However, these mostly laboratory tools are not suitable for the real-time online evaluation of learner progress during instruction. On the other hand, traditional educational tests either have limited diagnostic capabilities or are too time consum-

I thank Lauren Willis for her assistance in collecting experimental data.

Correspondence concerning this article should be addressed to Slava Kalyuga, School of Education, University of New South Wales, Sydney, New South Wales, 2052, Australia. E-mail: S.kalyuga@unsw.edu.au



ing. For example, common multiple-choice items at best indicate remote results of various possible solution strategies rather than information about actual solution steps. In such tests, knowledge-based expert solutions and trial-and-error novice efforts may receive the same scores. The rapid diagnostic approach attempts to build on the strengths of verbal reports and avoid the limitations of multiple-choice measures.

This article describes a rapid online diagnostic technique, the rapid verification method, and a series of experiments designed to validate the method in two different task domains. With this method, learners are required to rapidly verify suggested steps at various stages of the solution procedure. The method realizes a general rapid diagnostic approach based on our knowledge of human cognitive architecture, especially interactions between working memory (WM) and long-term memory (LTM) structures (Kalyuga, 2006b; Kalyuga & Sweller, 2004).

### A Rapid Testing Approach in Cognitive Diagnosis

WM is a setting for people's conscious cognitive processes. It is severely limited in duration and capacity when dealing with unfamiliar information (Baddeley, 1997; Miller, 1956; Peterson & Peterson, 1959). However, in well-known domains, the available knowledge base in LTM allows people to encapsulate many elements of information into larger chunks that are treated as single units in WM. Available LTM knowledge structures essentially define the characteristics of WM by effectively extending its capacity and reducing WM load. Therefore, WM capacity is not a stable characteristic of a learner but always depends on the learner knowledge base and directly reflects the available domain-specific knowledge structures. Accordingly, a diagnostic method based on immediate evaluation of units of WM content when an individual approaches a task could be used for measuring levels of expertise in the corresponding task domain. The method would actually evaluate the extent to which WM limits have been altered by domain-specific knowledge structures held in LTM.

Because of the association with stable LTM structures, the content of experts' knowledge units in WM is sufficiently durable and resistant to temporary interferences (Ericsson & Kintsch, 1995) to allow a working diagnostic procedure. If a person has a well-learned strategy for solving a specific class of tasks, he or she will immediately apply it when encountered with a task from this class. Someone without such knowledge will start searching for a solution randomly. It is practically possible to determine the content of knowledge (if any) used in a specific task situation, for example, by analyzing the concurrent (think-aloud) verbal reports. This method, however, is time consuming and difficult to use in online learning environments.

Some experimental procedures used in classical cognitive studies of chess expertise by De Groot (1965) and Chase and Simon (1973) suggested another idea. In those studies, professional grand masters performed considerably better than weekend players in reproducing briefly presented (for few seconds) chess positions taken from real games. There were no significant differences, however, for random configurations of chess figures. A schematic knowledge base of a huge number of different real game configurations held in grand masters' LTM allowed them to reproduce chess positions by familiar chunks rather than by individual chess figures without overloading WM.

Chess grand masters know the best move for each of the remembered typical real game configurations. Therefore, a rapid test of chess expertise could be based on presenting typical game configurations for brief periods of time and asking players to rapidly indicate their first move for each situation (essentially these are tasks that grand masters face when playing simultaneously on many chess boards). Applying this idea to the rapid diagnostic assessment of task-specific expertise in school mathematics resulted in the first-step diagnostic method (Kalyuga & Sweller, 2004). Learners were presented for a limited time with a series of tasks with gradually changing levels of complexity and were asked to rapidly indicate their first step toward solution of each task. The first step would involve different cognitive processes for individuals with different levels of task-specific expertise. For example, experts would bring their well-automated higher level procedures that allow them to rapidly generate advanced stages of the solution or even final answers to relatively more difficult tasks and skip many intermediate operations. On the other hand, complete novices may only be able to start the first attempt in their random search (e.g., trial-and-error) approach to even simple tasks. Therefore, different first steps would indicate not only the availability but also different levels of acquisition of corresponding knowledge structures in the learner's LTM.

Validation studies of the first-step method in algebra, coordinate geometry, and arithmetic word problem solving indicated significant correlations (.85, .92, and .72, respectively) between performance on these tasks and traditional measures of knowledge that required complete solutions of corresponding tasks (Kalyuga, 2006c; Kalyuga & Sweller, 2004). Test times were reduced by factors of up to 4.9 in comparison with traditional test times. The first-step method was sensitive to underlying knowledge structures and was sufficiently rapid. The method was used for optimizing levels of instructional guidance in adaptive computer-based tutoring in solving linear algebra equations (Kalyuga & Sweller, 2004, 2005). At the beginning of a session, each learner was allocated to an appropriate level of guidance according to the outcome of the initial rapid diagnostic test. Depending on the outcomes of the rapid diagnostic probes during instruction, the learner was allowed to proceed to the next stage of the session or was required to repeat the same stage and then take the rapid test again. At each subsequent stage of the tutoring session, a lower level of instructional guidance was provided to learners, and a higher level of the rapid diagnostic tasks was used at the end of the stage.

### Rapid Verification Method

Another possible approach to a rapid evaluation of chess expertise may be based on presenting a real game configuration for a brief period of time, followed by displays of several possible (both suitable and unsuitable) moves for this configuration, one display at a time. A player should rapidly verify the suitability of each of these moves. A similar approach to the rapid diagnostic assessment of task-specific expertise could also be applied in an online educational context. Learners could be presented with a series of possible (correct and incorrect) steps corresponding to various stages of the task solution procedure and asked to rapidly verify the suggested steps. Specific forms of responses could be clicking on displayed on-screen buttons or pressing specified keys on the computer keyboard (e.g., *correct*, *incorrect*, or *don't know*). An



important advantage of this rapid verification method over the first-step technique is its better suitability for computer-based and online applications, especially in task areas in which solution steps cannot be specified precisely in advance (e.g., when learners need to draw graphical task representations).

The ability to rapidly verify advanced stages of a solution procedure is based on the effectively expanded WM capacity and reduced cognitive load due to available task-relevant knowledge structures (e.g., procedures, rules). The levels of acquisition of this knowledge could be measured by an appropriately defined scoring method that reflects different levels of performance in verifying suggested solution steps. The required rapidness of learners' responses is not only a means of reducing testing time. More importantly, it is essential for capturing knowledge structures that learners actually use while approaching a task and before any lengthy chains of reasoning or searching could be applied, thus diagnosing the level of knowledge-based expertise.

The rapid diagnostic approach could be considered as a form of dynamic assessment that measures students' ability to move on with a task solution given a certain level of additional guidance (Bransford & Schwartz, 1999). In the case of rapid verification tasks, the additional guidance is provided by the depiction of a gradually changing number of previously completed steps. When presenting learners with various stages of a solution procedure for rapid verification, researchers use the varying levels of scaffolding to determine the level of learner proficiency in handling increasingly difficult situations.

The rapid verification method was actually used for optimizing levels of instructional guidance and individualizing instructional procedures in an adaptive computer-based tutorial in kinematics (Kalyuga, 2006a). As levels of learner expertise increased according to online rapid verification tests, the levels of provided guidance were reduced. On the other hand, for learners with insufficient levels of knowledge, more guidance was provided in the form of worked-out solution steps. However, in contrast to the first-step method, no validation studies were conducted for this method. Test scores are valid if they are obtained from a measure designed, or method implemented, to evaluate an attribute, construct, or variable that has been defined clearly on the basis of a theory and substantive reviews of literature. Using correlations of the test scores with other measures of the attribute as evidence that the test assesses the intended attribute is referred to as establishing the concurrent validity of the test. Thus, investigating whether the rapid assessment results would correlate highly with more traditional measures of task-specific expertise is needed to support the claim that the method actually measures the levels of learner expertise.

For example, there could be concerns that replacing the first-step rapid diagnostic method with a rapid verification technique would effectively turn a test of organized knowledge structures into a recognition test measuring knowledge of shallow task characteristics. The rapid verification method actually uses a recognition test format for verifying suggested solution steps. However, learners need to recognize intermediate advanced steps in a solution procedure, and these steps have to be rapidly constructed first by retrieving and using available knowledge in LTM. These processes involve more complex cognitive activities and higher levels of knowledge than those required by traditional recognition tests. Therefore, it is assumed that this method is diagnostically more

powerful than simple recognition tests and could be used for measuring levels of learner task-specific expertise. However, this assumption needs to be supported by studies of concurrent validity of the diagnostic procedure.

Thus, the experimental studies reported in this article were designed to answer the following research questions:

1. Is the suggested online verification procedure a valid diagnostic method capable of detecting different levels of acquisition of domain-specific knowledge structures? The hypothesis is that the rapid verification test results will correlate significantly with alternative traditional measures of task-specific expertise, thus demonstrating a high degree of concurrent validity.
2. Is the suggested online verification procedure a rapid method that could be completed fast enough for real-time applications? The hypothesis is that the rapid verification test will allow a substantial reduction of diagnostic assessment time in comparison with alternative traditional measures of task-specific expertise.
3. Will the suggested rapid verification procedure generalize to different task domains? The hypothesis is that the rapid tests will demonstrate high degrees of validity indicators in different task areas.

In addition to validating a new diagnostic method that implements the rapid diagnostic approach, the novelty of this study is also in using task domains (vector addition problems and graph transforming tasks) that essentially rely on graphical representations of task situations. Most previous studies of the rapid first-step diagnostic technique used numerical-only tasks. Another essential improvement is observing participants' actual problem-solving steps using video recordings and concurrent verbal reports as better criterion measures of task-specific expertise than students' records of problem-solving steps used in previous studies. Thus, in the traditional paper-based test, participants were required to provide complete solutions of tasks similar to those used in the rapid verification test. The participants' on-paper actions and think-aloud verbalizations were recorded and analyzed. In order to determine actual time reductions associated with rapid testing in comparison with the traditional test, I used self-paced tasks in both tests.

In both test conditions, more knowledgeable learners were expected to perform their tasks with lower levels of cognitive load than relative novices because of effectively increased WM capacity due to the available knowledge base. Therefore, I also included the evaluation of cognitive load in the procedure to provide another indicator of levels of learner expertise in addition to the task performance scores. Previous cognitive load research studies have indicated that a simple subjective rating scale of task difficulty could be an effective means of measuring cognitive load imposed by instructional materials (e.g., see Paas, Tuovinen, Tabbers, & van Gerven, 2003, for an overview).

If the tests actually measure levels of learner task-specific expertise, difficulty ratings are expected to show significant negative correlations with test scores: Because of reduced WM load, tasks in both tests should be relatively easier for more proficient learners than for less experienced participants. In this study, the measures of cognitive



load may also provide an indicator of participants' actual engagement in knowledge-based higher level cognitive activities during rapid assessment tasks. If the rapid tests were based only on simple recognition of surface task characteristics, the task difficulty would not differ for learners with different levels of expertise. Consequently, there would be no correlations between difficulty ratings and test scores for rapid verification tasks.

### Experiment 1

Experiment 1 was designed to investigate whether rapid verification scores in a task domain in kinematics could be validated by correlating highly with the results of observations of participants' paper-based problem solving steps and their concurrent verbal reports. The experimental tasks represented a class of problems in kinematics called *vector addition motion problems*. A typical task in this domain requires adding two vectors that are positioned at a certain angle to each other, for example, "A sea wave is traveling at 8 m/s towards the beach. A swimmer moves at 3 m/s in a direction perpendicular to the direction of the wave. What is the velocity of the swimmer relative to the ground?" During the rapid verification test, students were presented with a set of possible (correct and incorrect) intermediate solution steps and were asked to rapidly verify the correctness of these steps. More knowledgeable learners presumably should be better able to rapidly recognize more advanced intermediate solution moves than less knowledgeable learners.

For example, a person who knows that a vector approach should be applied but who has not practiced graphical addition of vectors may be able to verify correctly only a diagram with two perpendicular vectors as a valid step toward the solution. An individual who has more experience with vectors may rapidly verify perpendicular vectors with numerical values assigned to the length of each vector. Another person who is familiar with the vector addition procedure may also verify immediately a diagram representing the graphical addition of these vectors. Someone with more experience in adding vectors might be able to rapidly verify a numerical expression for the Pythagorean theorem that is used in solving this class of tasks. A learner with substantial experience in solving such tasks may even be able to verify a numeric expression representing the final answer without a diagram present.

### Method

#### Participants

Thirty-three university students (18 women and 15 men, age 18–25) participated in this experiment. They represented different years of study: 17 undergraduate, 3 graduate, and 13 postgraduate students. Participants also represented variety of subject areas: 12 students were involved in technical areas (including mathematics, mechanical engineering, computer engineering, biotechnology), and 21 were from nontechnical areas (education, psychology, medicine, law, management, international studies). In order to have a wide range of students' familiarity with the domain and different levels of their task-specific expertise, I deliberately recruited the participants from various areas of academic specializations with different degrees of involvement of mathematics. Brief pretest interviews indicated that participants represented different levels of expertise in the area of kinematics ranging from novices

(some still remembered that they had studied related material in their high school science courses) to experienced individuals studying university engineering and mathematics courses. The scores obtained from the rapid verification tasks and traditional measures of expertise in the experiment confirmed the initial pretest rough evaluations. Students with more expected experience in the domain performed better on the experimental tasks. For example, according to the values of rapid test scores (out of the maximum possible total score of 75), 2 participants representing nontechnical fields were in the range from 0 to 25 (the lower third), another 19 nontechnical participants were in the range from 26 to 50 (the medium third), and 11 participants representing technical areas were in the range from 51 to 75 (the upper third). The students had not been exposed to the specific materials used in the study prior to the experiment. They were paid AU\$20 for their participation in the experiment.

#### Materials and Procedure

Each participant was tested individually in a laboratory environment. Computer-based test items (designed using Authorware Professional; Macromedia, 2003) were delivered through a laptop personal computer. The experimental procedure included a rapid computer-based diagnostic test and a paper-based test with recording of students' on-paper actions and think-aloud verbal reports. The sequence of the test administration was counterbalanced: Approximately one-half of students performed the rapid test first, and the rest performed the paper-based test first.

**Rapid diagnostic test.** The test included five tasks corresponding to the following values of angles between vectors: 0° (the same direction of movements), 180° (opposite directions of movements), 90° (perpendicular vectors), 120°, and 60°. In addition, when 60° or 120° angles were used, only equal velocity values for both vectors were allowed. Without these restrictions, the procedures for calculating the length of the resulting vectors required more advanced knowledge of trigonometry that was not a part of the assessment objective in this study. Each textual task statement was followed by five suggested solution steps (correct or incorrect) for rapid verification.

The first verification subtask for each task provided vector graphs indicating only directions of movements. The second verification subtask provided vector graphs with velocity values indicated next to them. For example, for the previously mentioned task that described a situation with perpendicular directions of movements, vector graphs indicating perpendicular directions of movements with corresponding velocity values were provided. The third verification subtask, in addition to the vectors and their values, graphically represented the vector addition operation. For example, for the fourth task ("A boat is traveling at 5 m/s. A passenger runs across the deck at 5 m/s in a direction of 120° relative to the direction of motion of the boat. What is the velocity of the passenger relative to the water?"), the third verification subtask is presented in Figure 1 (incorrect step). The fourth verification subtask provided all necessary graphical information and indicated a numerical expression for calculating the length of the resulting vector. For example, for the above (120° angle) task, a simple expression  $V = 5 \text{ m/s}$  was placed next to the diagram (60° angles and equal sides in two equilateral triangles were also indicated on the diagram). Finally, the fifth verification subtask



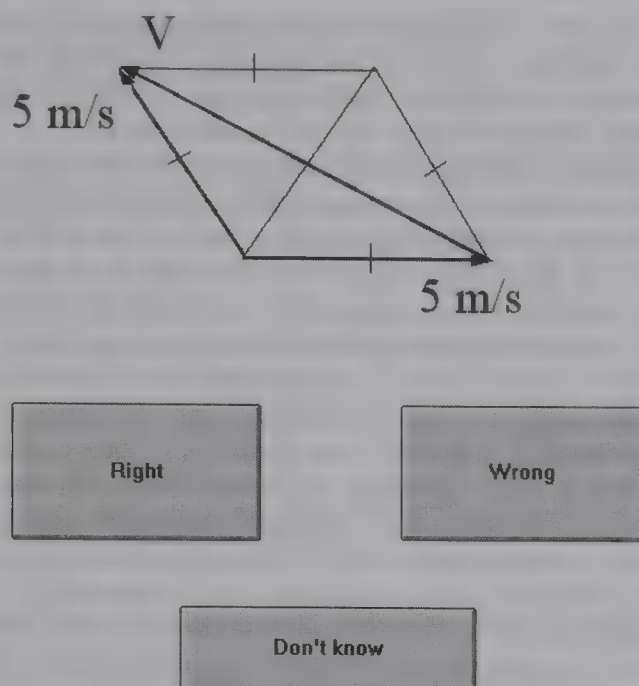


Figure 1. Snapshot of the rapid response window for a vector (V) addition motion task.

included final numerical answer (an integer or surd) with no graphics provided.

Students were instructed that each task in the test would be displayed for a limited time and that following each task, several possible (correct and incorrect) solution steps would be presented one at a time. The task exposure time (15 s) was established in preexperimental trials as sufficient for reading and comprehending task statements. Students were told that most of the suggested steps were possible intermediate stages on a way to the solution, but some suggested steps could also indicate final answers. For each suggested step, students had to immediately verify whether this could be a correct step leading to the solution (or providing the final answer). Each solution verification window included a diagrammatic and/or numerical representation of a possible solution step and the buttons *right*, *wrong*, and *don't know* for students to click on.

Before the rapid test, the participants were coached in responding sufficiently rapidly using exercises with simple tasks from a different area (common arithmetic calculations). During those exercises, the students got a sense of what was considered a rapid response. If a student did not respond within a set short time interval of few seconds, she or he was asked to respond faster next time and was recommended to try another exercise task. This brief procedural pretraining session was chosen instead of mechanically limiting the allowed verification response time to several seconds. The system-controlled response time (by automatically switching to the next verification window or task after a fixed short time interval) could forcefully interrupt genuine verification responses, thus invalidating results.

Each rapid diagnostic task was followed by a subjective rating of task difficulty, with a 9-point scale ranging from 1 (*extremely easy*) to 9 (*extremely difficult*) presented on the screen for students to click on. Students' response times, performance scores, difficulty ratings, and test times were automatically recorded by the software. In this experiment, the scores allocated for correct responses depended on the level of advancement of intermediate

solution stages represented in different verification subtasks. For example, the first subtask required learners to verify the application of only one step (a graphical representation of vectors), and a score of 1 was allocated for a correct response. On the other hand, the fifth subtask required learners to verify the result of the application of five sequential procedural steps, and a score of 5 was allocated for a correct response. Null scores were allocated for incorrect responses and "don't know" entries. Prior to the test, participants were instructed to enter "don't know" instead of guessing their responses any time they were in doubt. Because this was not a high-stake test, the students were expected to follow these instructions. Therefore, both incorrect responses and "don't know" entries were assumed to reflect a lack of specific knowledge, and null scores were allocated for both types of responses. The maximum possible total score for all five tasks in this test could be  $5 \times (5 + 4 + 3 + 2 + 1) = 75$ .

**Paper-based test.** Five textual task statements presented to learners were similar to those used in the rapid test (wording and numerical values in the task statements were changed). Each task statement was printed at the top of a separate page. Students were instructed to provide a complete solution for each task below the task statement as quickly as they could and think aloud while solving the task. Before commencing the procedure, participants were briefly coached in how to deliver a think-aloud protocol. The moderator (a research assistant) instructed them to think out loud at all times (e.g., "It really helps me to understand what you are thinking while solving the problem. If you get quiet, I will ask you to keep talking"). The moderator prompted participants to continue thinking aloud ("Please keep talking") every time they were silent for more than several seconds. Each diagnostic task was followed by a paper version of the subjective rating scale of task difficulty. The scale was identical to that used in the computer-based test; however, instead of clicking on their responses, students were asked to tick or circle them.

Time taken to complete the test was recorded for each participant. Students' problem-solving steps on paper as well as their verbalizations were recorded using a digital video camera. The camera was mounted on a tripod and focused on paper pages. Only the pages and students' hands were visually recorded. The research assistant collected the audiovisual performance records for each participant, and I subsequently analyzed them in order to evaluate levels of student task-specific expertise. The most important performance indicators that I looked for during this analysis were the following: search processes (if any) performed prior to executing specific solution steps, time taken by the participant before starting to carry out specific solution steps, and consistency and continuity of the solution process. For example, if the records demonstrated that a participant spent time on searching for possible moves at the beginning and/or subsequent stages of the solution process (rather than immediately applying available knowledge of relevant solution steps), her or his perceived level of expertise in this task domain was accordingly lowered.

A student's performance in each task was quantitatively assessed as the number of correct solution steps that the student had completed continuously within a short period of time (usually 10–20 s) of starting the solution. This score was determined on the basis of the analysis of both visual and audio recordings of the student's actions. The steps that were preceded by long chains of reasoning or searching and required more time did not count (even if they were eventually



completed correctly), because they were not based on immediately available knowledge of solution procedures in LTM. Thus, because five major steps were required to complete each of the five tasks, scores from 0 to 5 were allocated for a task. A total score out of 25 was assigned to each participant. The validity of this scoring method was supported by evident associations between the scores and expected levels of expertise based on participants' academic specializations: Students studying mathematics and technical subjects generally scored higher than those in nontechnical areas. For example, 11 out of 12 participants representing technical areas and only 5 out of 21 representing nontechnical areas were in the upper half (with scores above the mean).

Results and Discussion

Data for 1 student were lost because of a software problem. The variables under analysis were as follows: paper-based test time (time in seconds each learner spent on reading statements and solving all five test tasks),  $M = 476.75$ ,  $SD = 313.75$ ; rapid computer-based test time (time each learner spent on reading statements and verifying suggested steps for all five test tasks),  $M = 147.97$ ,  $SD = 20.33$ ; test scores for the traditional test,  $M = 13.12$ ,  $SD = 7.12$  (58% correct, actual range of test performance scores from 0 to 25); test scores for the rapid test,  $M = 46.19$ ,  $SD = 11.32$  (62% correct, actual range of test performance scores from 22 to 67); difficulty ratings averaged over all five tasks for the traditional test,  $M = 5.54$ ,  $SD = 2.00$  (actual range of ratings from 1.4 to 9.0; the rating 1 corresponded to extremely easy tasks); and average difficulty ratings for the rapid test,  $M = 4.68$ ,  $SD = 1.82$  (actual range of ratings from 1.0 to 8.6). Correlations between all variables are presented in Table 1.

A Pearson product-moment correlation,  $r(31) = .71$ ,  $p < .01$ , between scores for the traditional and rapid tests was obtained, with 95% confidence interval extending from 0.48 to 0.85, suggesting a high degree of the concurrent validity for the rapid test scores. A Pearson product-moment correlation between average ratings of task difficulty for the traditional and rapid tests was  $r(31) = .67$ ,  $p < .01$ , with 95% confidence interval extending from 0.42 to 0.83. There was a significant difference between test times for the traditional and rapid tests,  $t(31) = 5.90$ ,  $p < .01$ . Test time for the rapid method was reduced by a factor of 3.22 in comparison with the time for the traditional test. The average response time (in seconds) for verification subtasks was  $M = 3.04$ ,  $SD = 0.81$ , indicating that students actually responded rapidly, as they had been instructed and coached prior to the test. The estimates of Cronbach's coefficient alpha (.97 for the rapid test and .81 for the

traditional test) provided evidence of internal consistency of the scores. The high values are not surprising considering that the tasks measured closely associated attributes (a set of highly related procedural skills in solving a specific class of tasks).

As expected, difficulty ratings in both tests indicated significant negative correlations with test scores. These correlations reflect the fact that more knowledgeable learners experience lower WM load and perceive the tasks as being easier. Test times in the rapid test did not correlate with test times in the traditional test, thus indicating a qualitatively different nature of the rapid test format. The rapid test does not appear to be just an abbreviated version of the traditional diagnostic test. Nevertheless, both tests measure the same construct as indicated by the highly significant correlation between their scores. The rapid test scores did not correlate significantly with rapid test times, indicating that higher scores were likely due to immediately available knowledge base rather than longer time spent on searching for the correct response.

A negative correlation between test times and scores was expected for the traditional test: More knowledgeable learners should obviously finish earlier than less experienced students. Contrary to the expectation, a positive (although nonsignificant) correlation of .19 was obtained. The examination of video records revealed that in kinematics, the computationally intensive and very time-consuming last stage of the solution was rarely reached (or almost immediately abandoned) by novices; however, it was painstakingly performed by more experienced students, thus accounting for extended solution time for these participants.

The results of this experiment indicate highly significant correlations between learners' performance scores and difficulty ratings on the rapid verification tasks and traditional measures of learners' task-specific expertise, thus demonstrating a high degree of concurrent validity of the scores obtained from the suggested method. The rapid test was also completed much quicker than the traditional test. To further validate the rapid verification method, I applied it to a different task area in the following experiment.

Experiment 2

This experiment was designed as a correlation study in the task domain of transforming graphs of linear and quadratic functions in mathematics. Two tasks asked students to transform a provided graph of the basic line  $y = x$  into graphs of more complex lines,  $y = -2x + 3$  and  $y = \frac{1}{3}x - 2$  (see Figure 2 for an example of a task's statement). The following two tasks asked students to transform a provided graph of the basic line  $y = x^2$  into graphs of more complex quadratic functions,  $y = -\frac{1}{3}x^2$  and  $y = 2(x - 2)^2$ . The tasks required application of two or three of the following operations: flipping a graph because of the minus sign in front of  $x$  or  $x^2$  (the negative slope), squeezing (expanding) a graph toward (from) the y-axis according to the value of a coefficient in front of  $x$  or  $x^2$  (more or less than 1), and horizontal/vertical shifting. The aim of this study was to demonstrate that the rapid verification test in this task domain could be completed rapidly but that the resulting scores would have a high degree of concurrent validity.

Method

Participants

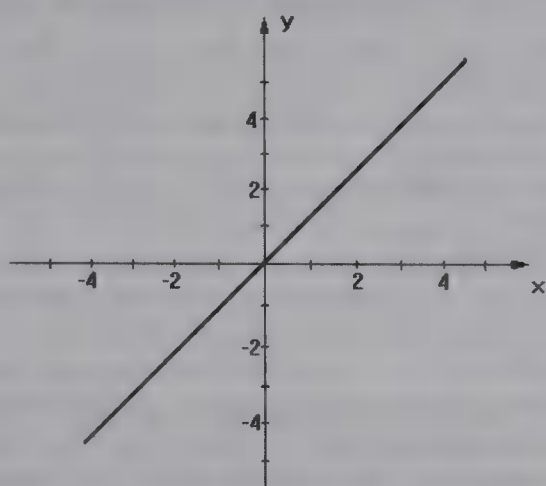
Thirty-three university students who took part in the previous study participated in this experiment. Because the task domain in

Table 1  
Correlations for Variables in Experiment 1

Variable	1	2	3	4	5	6
1. Rapid test scores	–	.71**	–.40*	–.53*	–.08	.34
2. Traditional test scores		–	–.48**	–.79**	–.11	.19
3. Rapid test difficulty			–	.67**	.24	–.33
4. Traditional test difficulty				–	.17	–.21
5. Rapid test time					–	–.05
6. Traditional test time						–

\* $p < .05$ . \*\* $p < .01$ .

This is a graph of the line  $y = x$ .



Transform it into a graph of the line  $y = -2x + 3$ .

Figure 2. Snapshot of the statement for a graph transformation task.

this study was completely different and unrelated to the previous experiment, using the same students could possibly not bias the results. On the contrary, it was expected that the participants' familiarity with the general format and demands of the rapid testing would eliminate its procedural novelty as a possible negative factor and increase the robustness and validity of results. In addition, the need for pretraining on the rapid testing procedure was also eliminated. In this sense, Experiment 1 could be regarded as a preliminary study in the different task domain, one of the purposes of which was to provide students with extensive pretraining in the rapid testing procedure in realistic conditions.

All students had been taught elementary mathematics in their high school courses. Such courses always include graphs of linear and quadratic functions. Some participants (especially those enrolled in nontechnical university courses) did not deal regularly with such problems and were relative novices in the task domain. Other participants (notably those studying mathematics and engineering courses) were more experienced in the domain. According to the values of total rapid test scores (the maximum possible score of 16), 3 participants were in the range from 0 to 4 (the bottom quarter), 14 participants were in the range from 5 to 8 (the second quarter), 14 participants were in the range from 9 to 12 (the third quarter), and 2 participants were in the range from 13 to 16 (the upper quarter). Students with more expected experience in the domain performed better on the experimental tasks (e.g., students studying mathematics and technical subjects generally scored higher than others in both tests). All participants had not previously encountered tasks formulated using the current format.

### Materials and Procedure

The experimental procedure was identical to that used in the previous experiment and included a rapid computer-based diagnostic test and a paper-based test with recordings of students' on-paper actions and concurrent verbal reports. The sequence of the test administration was counterbalanced.

**Rapid diagnostic test.** Each task statement was presented for 10 s (this time was established in preexperimental trials as suffi-

cient for comprehending very brief task statements) and was followed by four suggested solution steps for rapid verification. Students were instructed that most of the suggested steps were possible intermediate moves on a way to the solution, but some suggestions could also indicate final answers. Figure 3 shows an example of an incorrect intermediate step for the task represented in Figure 2. Some verification subtasks indicated results of the application of only one operation, whereas other subtasks indicated results of the application of several operations (e.g., flipping and expanding in Figure 3).

Because of the nature of the task, the scoring procedure in this study was different from the cumulative scoring approach used in the previous experiment. For the vector addition tasks, verification subtasks for each solution stage, except the final numerical answer, showed explicitly the fixed cumulative sequence of all prior steps that students would normally perform. For example, a diagram representing the graphical addition of vectors would necessarily show the vectors themselves with assigned numerical values (the essential attributes of the previous solution steps). In contrast, for the tasks used in this experiment, verification subtasks showed only results of the application of an ordered set of possible prior steps. An individual student might not necessarily solve the task using this specific sequence of steps. For example, when constructing a graph of the line  $y = -2x + 3$ , one person could first flip the line  $y = x$ , then squeeze the flipped line, followed by the shifting of the squeezed line up. Another individual would prefer to squeeze the original line first, then shift the squeezed line, and finish by flipping the shifted line. Yet another student would shift first and then perform one of the two possible sequences of the remaining steps.

As a result, in this task domain, the verification process is likely to be performed by locating a feature that would immediately

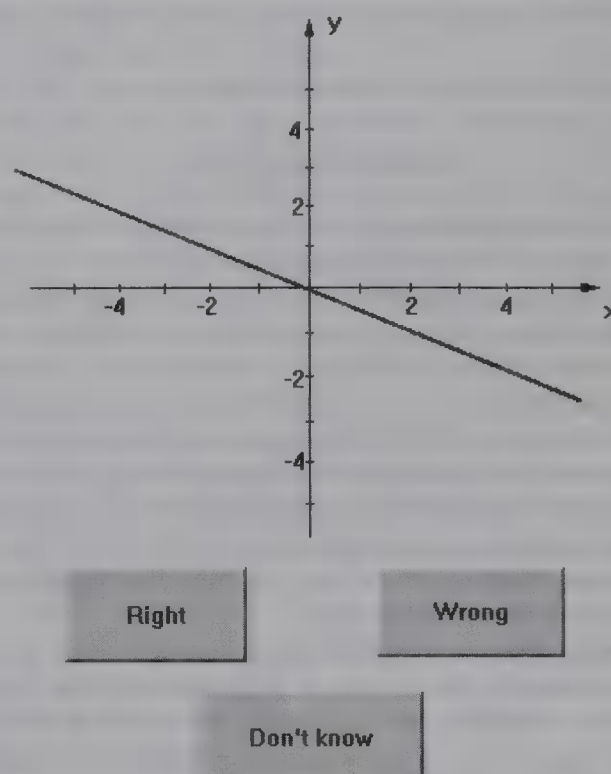


Figure 3. Snapshot of the rapid response window for a graph transformation task.



exclude the suggested step from a list of possible correct steps rather than by comparing the suggested step with different progressive stages of a mentally constructed solution sequence. For example, noticing that a flipped line is depicted for a function with a positive slope, or that an expanded line is depicted when the squeezing operation is required, or that a shift is made in the wrong direction would immediately flag an incorrect step. Because locating a single incorrect operation could be sufficient for the verification purpose, the scoring procedure in this task domain allocated a score of 1 for each correctly performed verification subtask. Therefore, the maximum possible score in the rapid test (4 tasks with 4 verification subtasks for each task) was  $4 \times 4 = 16$ . Each diagnostic task was followed by a 9-point subjective rating scale of task difficulty. Students' performance scores, difficulty ratings, and test times were automatically recorded by the software.

*Paper-based test.* Task statements presented to learners were similar to those used in the rapid test (numerical values in the task statements were changed). Students were asked to provide a complete solution for each task and think aloud while solving the task. A blank coordinate plane was provided on each of the four pages underneath the task statement. Students' problem-solving steps on paper and their concurrent verbalizations were recorded using a digital video camera. Each diagnostic task was followed by a paper version of the subjective rating scale. A student's performance on each task was scored as the number of correct steps that the student completed continuously within a short period of time (10–20 s) of starting the solution, based on the analysis of both visual and audio recordings of her or his actions. A total score out of 9 was allocated to each participant (for one task, the range of scores was from 0 to 3; for the remaining three tasks, the ranges were from 0 to 2). The adequacy of this scoring method was backed by clear associations between the scores and expected levels of expertise based on participants' academic specializations (students studying nontechnical subjects generally scored lower than those studying technical domains). For example, all 12 participants representing technical areas and only 3 out of 21 representing nontechnical areas were in the upper half (with scores above the mean).

### Results and Discussion

The variables under analysis were paper-based test time (time in seconds each learner spent on solving all four test tasks),  $M = 371.27$ ,  $SD = 193.71$ ; rapid computer-based test time (time each learner spent on solving all four test tasks),  $M = 105.52$ ,  $SD = 19.99$ ; test scores for the traditional test,  $M = 3.30$ ,  $SD = 2.85$  (37% correct, actual range of test performance scores from 0 to 9); test scores for the rapid test,  $M = 8.33$ ,  $SD = 2.77$  (52% correct, actual range of test performance scores from 2 to 15); average difficulty ratings for the traditional test,  $M = 3.63$ ,  $SD = 1.82$  (actual range of ratings from 1 to 7); and average difficulty ratings for the rapid test,  $M = 4.05$ ,  $SD = 1.83$  (actual range of ratings from 1 to 7.75). Correlations between all variables are presented in Table 2.

A Pearson product-moment correlation,  $r(32) = .75$ ,  $p < .01$ , between scores for the traditional and rapid tests was obtained, with 95% confidence interval extending from 0.55 to 0.87, suggesting a high degree of the concurrent validity for the rapid test scores. A Pearson product-moment correlation between average ratings of task difficulty for the traditional and rapid tests was  $r(32) = .82$ ,  $p < .01$ , with 95% confidence interval extending from

Table 2.  
*Correlations for Variables in Experiment 2*

Variable	1	2	3	4	5	6
1. Rapid test scores	–	.75**	–.55**	–.53**	–.23	–.40*
2. Traditional test scores		–	–.39*	–.47**	–.33	–.45**
3. Rapid test difficulty			–	.82**	.41*	.12
4. Traditional test difficulty				–	.43*	.22
5. Rapid test time					–	.16
6. Traditional test time						–

\* $p < .05$ . \*\* $p < .01$ .

0.66 to 0.91. There was a significant difference between test times for the traditional and rapid tests,  $t(32) = 7.97$ ,  $p < .01$ . Test time for the rapid method was reduced by a factor of 3.52 in comparison with the time for the traditional test. It should be noted that when the results of the rapid test in this experiment were rescored using a cumulative scoring procedure, similar to that used in Experiment 1, a nonsignificant Pearson product-moment correlation,  $r(32) = .21$ ,  $p = .25$ , was obtained between scores for the traditional and rapid tests. This significantly lower correlation value than that obtained using a simple scoring method indicates that the selection of an adequate scoring method is essential.

As in Experiment 1, there were significant negative correlations between difficulty ratings and test scores in both tests. These correlations reflect decreased WM load with increased levels of learner expertise. Test times in the rapid test did not correlate with test times in the traditional test, thus indicating a different nature of the rapid test format. As expected, rapid test scores did not correlate with rapid test times, and traditional test scores negatively correlated with traditional test times (less knowledgeable learners required more time to complete the test). The estimates of Cronbach's coefficient alpha (.96 for the rapid test scores and .85 for the traditional test scores) provided evidence that the reliability (internal consistency) of the test scores was very high.

Thus, significant correlations between performance scores as well as difficulty ratings obtained using rapid test tasks and traditional measures of expertise in this experiment demonstrated a high degree of concurrent validity of the rapid verification scores. The rapid test was also completed significantly quicker than the traditional test.

### General Discussion

Rapidly measuring levels of learner task-specific expertise is required for adapting instructional techniques and formats to changing levels of learner proficiency dynamically, in real time. The experiments described in this article were aimed at validating a rapid verification method for diagnosing levels of learner task-specific expertise. Students were presented with a series of possible (correct and incorrect) intermediate steps of a task solution and were asked to rapidly verify each step (i.e., to establish whether the suggested step was correct). Two experiments were designed to evaluate the concurrent validity of the rapid verification scores in two different task domains. Experimental data obtained from university students in task domains of vector addition problems in kinematics and graph transformation tasks in mathematics indicated significant correlations (correspondingly, .71 and .75) be-



tween performance scores on rapid verification tests and measures of expertise that were based on observations of students' detailed solutions of similar tasks. Rapid test times were reduced by factors of 3.2 (for kinematics tasks) and 3.5 (for mathematics tasks) in comparison with traditional test times. Thus, the results indicate that suggested tests could be completed rapidly and that their scores have a high degree of concurrent validity.

Values of Cronbach's coefficient alpha (.97 and .96) provided evidence that the reliability (internal consistency) of the rapid test scores in both experiments was sufficiently high. Task difficulty ratings and test times were used to obtain additional indications of validity of the rapid tests. As expected, task difficulty ratings correlated negatively with test scores in both experiments: Tasks were relatively easier for more expert learners (as measured by the suggested rapid test) than for less experienced participants. In rapid tests in both experiments, there were no correlations between test times and scores, indicating that participants' responses were likely based on their immediately available knowledge structures rather than on time-dependent search processes.

In both experiments, test times in the rapid tests did not correlate with test times in the corresponding traditional tests. This result provides an additional indication that rapid verification tests represent a qualitatively different diagnostic assessment format and not just abridged versions of the corresponding traditional tests. It could be assumed that rapid verification tasks involved a different set of cognitive processes than traditional tests (e.g., searching for a suggested verification solution step in the mentally constructed solution procedure for a task or searching for a feature that would immediately exclude a suggested step from possible correct steps).

Participants in this study ranged from undergraduate to postgraduate university students. As noted by one of the reviewers, the assumption was that each student can be ranked with respect to expertise. There could be a potential confound due to different student abilities: Individuals who finish a graduate school may be brighter than undergraduates in general. However, according to one of the cognitive sciences most important and established results, the level of expertise associated with the available knowledge base is the most important factor influencing student learning and performance, and in most situations, it would override other relevant factors.

Thus, in response to the three research questions formulated for this empirical study, it is possible to state that (a) the suggested online verification procedure is a valid diagnostic method capable of detecting different levels of acquisition of domain-specific knowledge structures, (b) the suggested online verification procedure is a rapid method that could be completed fast enough for real-time applications, and (c) the suggested rapid verification procedure is likely to generalize to different task domains. The rapid verification method, as well as the previously investigated first-step method (Kalyuga & Sweller, 2004), are capable of providing evidence about learner levels of expertise that is diagnostically more valuable and powerful than that obtained in traditional (e.g., multiple-choice) tests. The value of this evidence may approach that of verbal reports, however, with testing times significantly reduced compared with alternative assessment methods.

Applying the rapid verification method to a task generally includes the following steps: (a) establish a sequence of possible main intermediate stages in the solution procedure for the task; (b) for each stage, select representative (either correct or incorrect)

solution steps; (c) present the original task statement to a learner for a limited time sufficient to read and understand the statement; (d) present a series of the selected intermediate solution steps to learners, one at a time, with the requirement to quickly verify whether each of the suggested steps could lead to a complete solution of the task; (e) select a scoring method depending on the uniqueness of the solution sequence for the task. If there is only one possible solution path, use a cumulative scoring method (allocate higher scores for more advanced steps that are verified correctly, with one unit added to the score for each level of advancement). If there is a range of possible solution paths, use a simple scoring method (a score of 1 is assigned for each correctly verified step).

For example, in the class of simple linear equations of the type  $ax + b = c$  (solve for  $x$ ), two main solution stages are subtracting  $b$  from both sides of the equation and dividing both sides of the resulting equation by  $a$ . Therefore, for the representative task  $3x + 2 = 5$  (solve for  $x$ ), the first step in the design of the rapid verification test is establishing sequential stages of the solution procedure:  $3x + 2 - 2 = 5 - 2$ ;  $3x = 3$ ;  $3x/3 = 3/3$ ; and  $x = 1$ . At the second step, representative solution steps could be selected, for example,  $3x + 2 - 5 = 5 - 5$  (incorrect);  $3x = 3$  (correct);  $3x/5 = 3/5$  (incorrect);  $x = 2$  (incorrect). Then, the original task statement should be presented to a learner for a few seconds followed by the selected representative solution steps, one at a time. The learner should quickly determine whether the suggested steps could lead to a complete solution of the task. Finally, because there is only one possible solution path in this case, a cumulative scoring procedure needs to be applied. Scores of 4 or 3 should be allocated, respectively, for verifying rapidly the suggested final answer or the step immediately preceding it (in the provided example,  $x = 2$  and  $3x/5 = 3/5$ , respectively). A score of 2 should be allocated for a correct answer at the stage of completed application of the first procedural step ( $3x = 3$ ) and a score of 1 for verifying an incomplete intermediate step in the application of the first step ( $3x + 2 - 5 = 5 - 5$ ).

This study has been limited to two task classes associated with sufficiently well-structured tasks and predictable solution paths. In such areas, different levels of expert behavior could be described as sequential solution stages a person can complete immediately on starting the solution process. More ambiguous and poorly specified task domains, for example, areas that involve problems with multiple possible routes to solutions, may require special content validation procedures in order to establish expertlike solution paths. The rapid verification method could possibly be used in such relatively poorly structured task domains. Only a limited number of situations or steps representing different possible stages of valid solution procedures could be selected and included into rapid verification subtasks. For example, for a medical diagnosis task, a sequence of progressively more advanced stages of testing different hypotheses (including both suitable and inappropriate steps) could be presented for rapid verification. Because of the variety of possible solution paths in this case, a simple scoring procedure should be applied. In further research, the generality and limits of usability of the method, especially in poorly structured task domains, need to be established.

In some areas, the rapid assessment approach (both the first-step and rapid verification methods) could be more suitable for measuring level of expertise of relatively advanced learners rather than



for thorough cognitive diagnosis of novices. Novice learners may have knowledge deficits of types that could not be anticipated in advance when selecting relevant possible solution steps for verification or programming the scoring engine (e.g., linguistic comprehension problems, insufficient factual knowledge, lack of basic metacognitive planning and monitoring skills). Most of these types of knowledge are usually taken for granted when dealing with relatively more experienced learners.

As a form of dynamic assessment, the method could be instrumental in developing adaptive expertise (Bransford et al., 1999) by enabling the selection of optimal environments for building flexible skills. On the basis of the fine-grained diagnosis of the acquired knowledge and skills in a task domain, such environments would provide learners with individually tailored instructional support that is optimal for moving on to the next level of expert performance. Acquisition of task-specific expertise is necessary for releasing cognitive resources for dealing with more nonroutine and creative aspects of expert performance. With learner-tailored, just-in-time forms of support, learners would be capable of efficiently handling relatively new situations without a cognitive overload and associated loss of interest in the task.

A preliminary study indicated that the suggested diagnostic method could be successfully used for the dynamic selection of appropriate levels of instructional guidance that are optimal for learners with different levels of task-specific expertise (Kalyuga, 2006a). The method was used in an adaptive tutorial in the domain of kinematics for online monitoring of current levels of learner expertise. At the beginning of a tutoring session, each student was allocated to an appropriate level of guidance according to the outcome of the initial rapid verification test. Depending on the outcomes of rapid verification probes during instruction, the learner was allowed to proceed to the next stage or required to repeat the same stage and then take the rapid test again. At each subsequent stage, a lower level of instructional guidance was provided to learners by gradually fading fully worked-out steps and increasing the number of steps that learners had to complete on their own. Also, a higher level of the rapid verification task was used at the end of the stage. The adaptive approach proved to be superior to traditional nonadapted instructional formats. In the future, more comprehensive studies are needed for comparing different dynamic adaptive methodologies based on rapid diagnostic techniques. Optimized adaptive learning environments need to be developed and tested in relatively more complex and less structured domains than tasks in mathematics and kinematics.

Researchers' knowledge of cognitive processes in learning and problem solving has advanced substantially, and many cognitively supported instructional methods have been developed. Efficient tailoring of these instructional methods to learners with different levels of task-specific expertise is dependent on researchers' ability to rapidly and accurately measure these levels of expertise in real time. This article is intended as a step in developing such rapid online diagnostic methods.

## References

Alexander, P. A. (2004). A model of domain learning: Reinterpreting expertise as a multidimensional, multistage process. In D. Y. Dai & R. J.

- Sternberg (Eds.), *Interactive models* (pp. 271–298). Mahwah, NJ: Lawrence Erlbaum Associates.
- Baddeley, A. (1997). *Human memory: Theory and practice*. East Sussex, England: Psychology Press.
- Blessing, S. B., & Anderson, J. R. (1996). How people learn to skip steps. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 576–598.
- Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (1999). *How people learn: Mind, brain, experience, and school*. Washington, DC: National Academy Press.
- Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of Research in Education*, 24 (pp. 61–101). Washington DC: American Educational Research Association.
- Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, 4, 55–81.
- De Groot, A. D. (1965). *Thought and choice in chess*. The Hague, The Netherlands: Mouton.
- Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist*, 49, 725–747.
- Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, 102, 211–245.
- Kalyuga, S. (2005). Prior knowledge principle in multimedia learning. In R. Mayer (Ed.), *Cambridge Handbook of Multimedia Learning* (pp. 325–337). New York: Cambridge University Press.
- Kalyuga, S. (2006a). Assessment of learners' organized knowledge structures in adaptive learning environments. *Applied Cognitive Psychology*, 20, 333–342.
- Kalyuga, S. (2006b). *Instructing and testing advanced learners: A cognitive load approach*. New York: Nova Science Publishers.
- Kalyuga, S. (2006c). Rapid assessment of learners' knowledge structures. *Learning & Instruction*, 16, 1–11.
- Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). Expertise Reversal Effect. *Educational Psychologist*, 38, 23–31.
- Kalyuga, S., & Sweller, J. (2004). Measuring knowledge to optimize cognitive load factors during instruction. *Journal of Educational Psychology*, 96, 558–568.
- Kalyuga, S., & Sweller, J. (2005). Rapid dynamic assessment of expertise to improve the efficiency of adaptive e-learning. *Educational Technology, Research and Development*, 53, 83–93.
- Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science*, 14, 511–550.
- Macromedia. (2003). *The Authorware Professional 7.0* [Computer software]. San Francisco, CA: Author.
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81–97.
- Paas, F., Tuovinen, J., Tabbers, H., & van Gerven, P. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, 38, 63–71.
- Peterson, L., & Peterson, M. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, 58, 193–198.
- Sweller, J., Mawer, R., & Ward, M. (1983). Development of expertise in mathematical problem solving. *Journal of Experimental Psychology: General*, 112, 639–661.
- Van Merriënboer, J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review*, 17, 147–177.

Received August 24, 2006

Revision received August 9, 2007

Accepted August 10, 2007 ■

# On the Measurement of Achievement Goals: Critique, Illustration, and Application

Andrew J. Elliot  
University of Rochester

Kou Murayama  
Tokyo Institute of Technology

The authors identified several specific problems with the measurement of achievement goals in the current literature and illustrated these problems, focusing primarily on A. J. Elliot and H. A. McGregor's (2001) Achievement Goal Questionnaire (AGQ). They attended to these problems by creating the AGQ-Revised and conducting a study that examined the measure's structural validity and predictive utility with 229 (76 male, 150 female, 3 unspecified) undergraduates. The hypothesized factor and dimensional structures of the measure were confirmed and shown to be superior to a host of alternatives. The predictions were nearly uniformly supported with regard to both the antecedents (need for achievement and fear of failure) and consequences (intrinsic motivation and exam performance) of the 4 achievement goals. In discussing their work, the authors highlight the importance and value of additional precision in the area of achievement goal measurement.

*Keywords:* goal, performance, mastery, approach, avoidance

For the past two decades, achievement goals have been a central construct in the study of motivation in achievement settings. Some achievement goal research has been experimental in nature, manipulating goals and examining their effect on outcomes relevant to achievement. However, the vast majority of achievement goal research has been correlational, measuring preexisting goals and examining the antecedents and consequences of these goals in concurrent, prospective, and occasionally longitudinal designs. Both the experimental and the correlational research have yielded a substantial amount of information about the strivings of individuals (most commonly students, athletes, and employees) in achievement contexts and the implications of these strivings (see Duda, 2005; Elliot, 2005; Meece, Anderman, & Anderman, 2006; Payne, Youngcourt, & Beaubien, 2007; Ryan, Ryan, Arbuthnot, & Samuels, 2007). Beyond doubt, the achievement goal construct represents a landmark contribution to the century-long study of competence and motivation.

Although clearly informative and generative, the achievement goal approach also faces its share of challenges and difficulties. Perhaps foremost among these challenges and difficulties is a long-term struggle to assess achievement goals in a conceptually rigorous manner. Some achievement goal measures rest on a weak

foundation in that the achievement goal concept is not clearly articulated a priori, thereby providing little guidance for how goals should be operationalized. Even when a clear conceptualization of achievement goals is in place, however, there is often poor correspondence between how the goals are conceptualized and how they are operationalized. This poor correspondence is of great consequence, because it makes it difficult to interpret empirical results straightforwardly and confidently, whether they are supportive or unsupportive of theoretical predictions. Interpretational ambiguity, in turn, retards theoretical progress in the achievement goal literature and undermines attempts to transfer information gleaned from research to real-world achievement settings.

In the present article, we identify several specific problems with the measurement of achievement goals in the current literature. We focus primarily on one achievement goal measure—Elliot and McGregor's (2001) Achievement Goal Questionnaire (AGQ)—to explicate and illustrate these problems. We begin by describing the conceptual foundation from which the AGQ emerged. We then show different ways in which particular AGQ items do not optimally correspond to this conceptual foundation and describe how such problems may be rectified. We also point to other examples in the achievement goal literature to demonstrate that these problems are not unique to the AGQ. The result of critiquing and adjusting the AGQ is a new achievement goal measure, the AGQ-Revised (AGQ-R), which we proceed to empirically test. Specifically, we examine the structural validity of the measure using both established and novel procedures and report data on the predictive utility of the goal subscales from this measure. Our aim is to demonstrate that the problems with existing achievement goal measures can be rectified while retaining (and perhaps, in some instances, even enhancing) the reliability and validity of the original measures.

## Conceptual Foundation

Elliot and McGregor's (2001) AGQ was designed to assess achievement goals as conceptualized in the  $2 \times 2$  achievement

---

Andrew J. Elliot, Department of Clinical and Social Psychology, University of Rochester; and Kou Murayama, Graduate School of Decision Science and Technology, Tokyo Institute of Technology, Tokyo, Japan.

This research was supported by a grant from the William T. Grant Foundation to Andrew J. Elliot. We extend thanks to the Approach-Avoidance Motivation lab at the University of Rochester for their assistance with data collection and their comments on various drafts of this article.

Correspondence concerning this article should be addressed to Andrew J. Elliot, Department of Clinical and Social Sciences in Psychology, University of Rochester, RC Box 270266, Rochester, NY 14627-0266. E-mail: andrew.elliott@rochester.edu



goal framework (Elliot, 1999; Elliot & McGregor, 2001) within the hierarchical model of approach–avoidance achievement motivation (Elliot, 1997, 2006). From this perspective, achievement goals are conceptualized as cognitive–dynamic aims that focus on competence, and any given achievement goal is thought to contain components from two independent competence dimensions. The *definition* dimension forms the basis of a mastery–performance distinction, which has been a part of the achievement goal tradition since its inception (Maehr & Nicholls, 1980). Competence may be defined in terms of the standard used to evaluate it, that is, relative to an absolute or intrapersonal standard (mastery) or relative to a normative standard (performance). Mastery-based standards tend to focus individuals on learning, whereas performance-based standards tend to focus individuals on performing (Dweck, 1986). The *valence* dimension of competence forms the basis of an approach–avoidance distinction, a later addition to the achievement goal tradition (Elliot & Harackiewicz, 1996). Competence may be valenced in terms of whether it is focused on a positive possibility to approach (i.e., success) or a negative possibility to avoid (i.e., failure). Combining the mastery–performance and approach–avoidance distinctions leads to four different types of achievement goals: mastery–approach (focused on attaining task-based or intrapersonal competence), performance–approach (focused on attaining normative competence), mastery–avoidance (focused on avoiding task-based or intrapersonal incompetence), and performance–avoidance (focused on avoiding normative incompetence).

When viewed from a hierarchical standpoint (see Elliot, 2006), the achievement goals of the  $2 \times 2$  framework are posited to emerge from more general motivations (e.g., the need for achievement and fear of failure), self-conceptions and theories (e.g., entity and incremental theories of ability), and environmental emphases (e.g., classroom goal structures). These goals are viewed as the proximal predictors of important achievement-relevant outcomes such as intrinsic motivation and performance attainment (Elliot & Church, 1997).

### Problems in Achievement Goal Measures

A host of achievement goal measures have appeared in the educational psychology, industrial–organizational psychology, social-personality psychology, and sport and exercise psychology disciplines over the past two decades, some of which have focused on the mastery–performance distinction alone, and others of which, like the AGQ, have focused on both the mastery–performance and approach–avoidance distinctions. In the following, we identify a number of problems that have appeared in these measures. We do so using the AGQ in illustrative fashion, as the focal point of both critique and solution.<sup>1</sup>

#### *Failing to Assess Goals*

In the  $2 \times 2$  framework, *goal* is conceptualized as an aim that one is committed to that serves as a guide for future behavior (Elliot, 1999; Elliot & Fryer, 2008). However, the prefixes of some AGQ items seem to suggest a value (e.g., “It is important for me to do better than other students”) or a concern (e.g., “I worry that I may not learn all that I possibly could in this class”), rather than a goal per se. Our solution to this problem is to select the same set of three prefixes for each goal scale in the AGQ-R, each of which

is exclusively goal-based (“My goal is to . . .,” “My aim is to . . .,” and “I am striving to . . .”).

Most achievement goals theorists would likely agree that achievement goals are best construed in terms of purposeful commitments that guide future behavior (Dweck & Elliott, 1983; Maehr, 1989). Nevertheless, taken at face value, many goal items do not seem to be temporally focused and do not appear to assess intentional commitments. Some items ask respondents to report on how they define success (e.g., “I feel most successful when . . .”; Button, Mathieu, & Zajac, 1996; Duda, Chi, Newton, Walling, & Catley, 1995; Duda & Nicholls, 1992; Midgley et al., 2000 [original]; Nicholls, Cobb, Wood, Yackel, & Patashnick, 1990; Nicholls, Patashnick, & Nolen, 1985; Roberts & Treasure, 1995; Roedel, Schraw, & Plake, 1994; Skaalvik, 1997). Others ask respondents to report on their values or concerns (e.g., “I value . . .”; Bouffard, Boisvert, Vezeau, & Larouche, 1995; Button et al., 1996; Conroy, Elliot, & Hofer, 2003; Elliot & Church, 1997; Harackiewicz, Barron, Elliot, Carter, & Lehto, 1997; Meece, Blumenfeld, & Hoyle, 1988; Middleton & Midgley, 1997; Midgley et al., 2000 [original and revised]; Roedel et al., 1994; Skaalvik, 1997; Stipek & Gralinski, 1996; VandeWalle, 1997; Zweig & Webster, 2004). Finally, others ask respondents to indicate what they like or prefer, or are pleased or satisfied by (e.g., “I feel really pleased when . . .”; Bouffard et al., 1995; Button et al., 1996; Elliot & Church, 1997; Harackiewicz et al., 1997; Middleton & Midgley, 1997; Midgley et al., 2000 [original]; Roedel et al., 1994; Skaalvik, 1997; Stipek & Gralinski, 1996; VandeWalle, 1997; Zweig & Webster, 2004). Although such items may capture goal adoption or lead to goal pursuit, more direct and precise wording is clearly both possible and desirable.

#### *Collapsing Together the Goal and the Motivation Underlying the Goal*

In the  $2 \times 2$  framework, a goal is construed as a cognitively represented aim, and this aim is viewed as separate from the reason or reasons why the person is pursuing the aim (Elliot & Thrash, 2001). As such, mastery-based and performance-based goals are differentiated with regard to the specific type of competence (incompetence) that one seeks to approach or avoid, and any additional reasons for such striving are excluded from the goal construct per se. From this standpoint, it is best to assess the reason behind the goal separately from the goal itself, thereby allowing the possibility of numerous achievement goal “complexes” (i.e., goal–reason combinations; see Elliot & Thrash, 2001, and Thrash & Elliot, 2001, for a discussion of the benefits of keeping goals conceptually separate from reasons). In the AGQ, however, one of the performance–avoidance goal items collapses the goal together

<sup>1</sup> Our review of other achievement goal measures is meant to be broad in scope but not comprehensive. The measures that we focus on herein are published in highly visible journals, have been widely used in the achievement goal literature, or contain items that are particularly fine examples of the points we seek to highlight. Some of these measures are currently used in achievement goal research, whereas others are no longer used (although data from these measures continue to be used to support conceptual points in the literature). It should be noted that not all measures that are cited together in the text within any given point necessarily share the same theoretical perspective.



with an underlying motive, fear of failure: "My fear of performing poorly in this class is often what motivates me"; indeed, it focuses more on the motive than the goal. Our solution to this problem is to omit the motive content in the AGQ-R version of this item, thus allowing fear of failure to be assessed separately and the link between fear of failure and performance-avoidance goals to be examined without item overlap.

Our view of goal as "aim separate from reason" is consistent with most conceptualizations of the goal construct in scientific psychology (Elliot & Fryer, 2008), but is different from the way that many, if not most, achievement goal theorists construe goals. Achievement goal theorists tend to prefer a combined aim-reason goal construct (Ames, 1992; Pintrich & Schunk, 1996); thus, items that combine aim and reason clearly would not be considered a problem for such theorists. Indeed, performance-based goal measures commonly include item content involving demonstrating or showing something to others, as well as a focus on normative competence (Button et al., 1996; Elliot & Church, 1997; Greene & Miller, 1996; Harackiewicz et al., 1997; Meece et al., 1988; Midgley et al., 2000 [revised]; Nicholls et al., 1985; Roberts & Treasure, 1995; Roedel et al., 1994; Skaalvik, 1997; Stipek & Gralinski, 1996). Some performance-based goal scales even focus on demonstration to the exclusion of normative competence (Middleton & Midgley, 1997; Midgley et al., 2000 [original]; VandeWalle, 1997; Zweig & Webster, 2004). Furthermore, in some measures, the performance-approach goal scale contains no or little mention of demonstration, whereas the performance-avoidance goal scale focuses exclusively or nearly exclusively on demonstration (Middleton & Midgley, 1997; Midgley et al., 2000 [original]; Skaalvik, 1997). In another measure, the performance-approach goal scale focuses exclusively on demonstration, whereas the performance-avoidance goal scale makes no mention of demonstration (Zweig & Webster, 2004). If reasons are accepted as part of a goal construct, it seems optimal for corresponding goal measures to balance the degree to which a focal reason, such as demonstration, is present in the two types of performance-based goals.

#### *Item Content Applicable to Both Mastery-Based and Performance-Based Goals*

In the  $2 \times 2$  framework, mastery-based and performance-based goals are differentiated in terms of their competence foci. In the AGQ, one performance-approach goal item contains content regarding grades: "My goal is to get a better grade than most of the other students." Grades can be applicable to either mastery-based or performance-based goals, depending on the nature of performance evaluation in the achievement setting (e.g., a task-based or normative grading structure). Our solution to this problem is to simply omit the reference to grades in this item in the AGQ-R.

Most achievement goal theorists would agree that mastery-based and performance-based goals focus on different types of competence (Dweck, 1986; Nicholls, 1989), but measures commonly contain content that seems applicable to both types of competence. For example, as with the aforementioned AGQ item, grades are commonly mentioned in performance-approach goal items (Bouffard et al., 1995; Dweck, 1999; Elliot & Church, 1997; Harackiewicz et al., 1997; Roedel et al., 1994). Mastery-approach goal items sometimes include content assessing the degree to

which the individual keeps busy (Nicholls et al., 1985, 1990), works hard (Duda et al., 1995; Duda & Nicholls, 1992; Nicholls et al., 1985, 1990; Roberts & Treasure, 1995), perseveres (Roedel et al., 1994), or reaches a goal (Roberts & Treasure, 1995); each of these characteristics is relevant to performance-based, as well as mastery-based, goals.

#### *Pitting One Goal Against Another*

In the  $2 \times 2$  framework, achievement goals are not presumed to be negatively correlated but instead are expected to be positively correlated (when they share a dimension) or uncorrelated (when they do not share a dimension). Accordingly, achievement goals of various types may be pursued at the same time, and it is best to assess each goal separately from the others. In the AGQ, however, one of the performance-avoidance goal items uses the word *just* ("I just want to avoid doing poorly in this class") to subtly imply the exclusion of other goals. Our solution to this problem is to simply omit this subtle reference to exclusivity in the AGQ-R items.

Some achievement goal theorists explicitly embrace the possibility of multiple goal pursuit (Barron & Harackiewicz, 2001; Pintrich, 2000), and even theorists who emphasize the possibility of performance-approach goals driving out mastery-approach goals do not portray these goals, or any others, as strongly negatively correlated in most achievement settings (see Midgley, Kaplan, & Middleton, 2001, p. 83). Nevertheless, several achievement goal measures include items that play off one goal against another. For example, in the item "Although I hate to admit it, I sometimes would rather do well in a class than learn a lot" (Dweck, 1999, p. 185) the respondent is asked to choose between a mastery-approach goal and an implicit performance-approach goal. In other measures, certain items put goals in competition with each other in a more subtle fashion (e.g., "I like it best when something I learn makes me want to find out more," Harackiewicz et al., 1997, p. 319; see also Elliot & Church, 1997; Middleton & Midgley, 1997; VandeWalle, 1997). There is certainly room for measures that focus on contrasting or rank ordering achievement goals, but it is important that these measures be explicitly identified as such and that they be used with full awareness of their limitations regarding multiple goal adoption (see Van Yperen, 2006, for an example of such a measure).

#### *Performance-Approach and Performance-Avoidance Goals That Differentially Emphasize Normative Comparison*

In the  $2 \times 2$  framework, normative comparison is an integral feature of performance-based goals. In the AGQ, the performance-approach goal items are explicitly normative in content, whereas the performance-avoidance goal items do not make specific mention of normative comparison (e.g., "My goal in this class is to avoid performing poorly"). Although it is likely that normative comparison would be implicitly read into most of these performance-avoidance goal items in most achievement contexts, it is clearly best to make the explicitness of normative comparison comparable across the two performance-based goals. As such, explicit normative content is added to each of the performance-avoidance goal items in the AGQ-R.



Most achievement goal theorists would concur that normative comparison and performance-based goals are closely connected (Maehr, 1983; Nicholls, 1984). Nevertheless, in some achievement goal measures, explicit normative content is either missing from performance-approach or performance-avoidance goals altogether (Elliot & Church, 1997; Middleton & Midgley, 1997; Midgley et al., 2000 [original]; VandeWalle, 1997) or there is an unequal proportion of normatively focused items in the two types of performance-based goals (Midgley et al., 2000 [revised]; Skaalvik, 1997; Zweig & Webster, 2004).

### *Performance-Approach and Performance-Avoidance Goals That Focus on Extreme Groups*

In the  $2 \times 2$  framework, performance-approach and performance-avoidance goals are presumed to be applicable across levels of perceived competence. Perceived competence is viewed as an important antecedent of these goals, but these relations are of moderate strength. As such, although high perceived competence tends to promote performance-approach goals, those with low and moderate perceived competence may also strive to do well relative to others; likewise, although low perceived competence tends to evoke performance-avoidance goals, those with high and moderate perceived competence may also strive to avoid doing poorly relative to others. It is possible for performance-based goal items to make high or low normative performance salient in a general way, without highlighting extremes (e.g., “do better than others” or “not do worse than others”); the more general the normative referent, the more likely the items will be to assess differences in valence per se, rather than differences in valence coupled with differences in perceived competence. In the AGQ, however, the word *most* in the performance-approach goal item, “My goal in this class is to get a better grade than most of the other students,” unnecessarily highlights a specific portion of the normative distribution. Our solution to this problem is to simply omit this qualifier in the AGQ-R item.

We are not aware of any achievement goal theorist who would espouse restricting the focus of performance-based goals to extreme groups. Nevertheless, in some achievement goal measures, performance-based goal items focus respondents on extremes in the normative distribution—specifically, being one of the best performers for performance-approach goals or not being one of the worst performers for performance-avoidance goals (Conroy et al., 2003; Elliot & Church, 1997; Middleton & Midgley, 1997; Midgley et al., 2000 [original]; Skaalvik, 1997).

### *Mastery-Approach/Mastery-Avoidance and Performance-Approach/Performance-Avoidance Goals That Contain Differential Amounts of Affective Content*

In the  $2 \times 2$  framework, affect is implied whenever any sort of goal commitment is present because this commitment represents an affective investment of the self with regard to a future possibility (Custers & Aarts, 2005; Elliot & Fryer, 2008). However, affect per se is not the central focus of the goal construct, and therefore it would be ideal if achievement goal items either were devoid of specific affective reference or, at minimum, distributed such references equally across the different types of goals. In the AGQ, the prefix of some of the mastery-avoidance goal items

contains specific affective content (e.g., “I’m afraid . . .”), whereas this is not the case for the mastery-approach goal items; likewise, one of the performance-avoidance goal items contains specific affective content (“My fear of performing poorly in this class is often what motivates me”), but none of the performance-approach goal items contain such content. Our solution to this problem is to omit explicit reference to affective content from the goal items altogether in the AGQ-R.

Some achievement goal theorists may be open to incorporating specific affective content in achievement goal items, but none would suggest that the amount of such content should vary across goals. Nevertheless, such variation is present in some measures for mastery-approach and mastery-avoidance goals (Conroy et al., 2003), as well as for performance-approach and performance-avoidance goals (Elliot & Church, 1997; Middleton & Midgley, 1997; Midgley et al., 2000 [original]; Skaalvik, 1997; Zweig & Webster, 2004).

Table 1 provides the items for the AGQ-R and pairs each revised item with the original AGQ item that it is meant to replace. This revised measure is designed to assess, in face valid fashion, the four goals of the  $2 \times 2$  achievement goal model while taking care to attend to each of the potential pitfalls that we have identified. We do not hold up the revised measure as the end of the measurement development process. Rather, we see it as an important step toward enhanced rigor and precision in an ongoing process. Furthermore, our critique of the AGQ and existing achievement goal measures is not meant to invalidate these measures or the empirical work that has been produced with them. Despite the aforementioned issues, we think that these measures have for the most part done a reasonably good job of assessing achievement goals and that the achievement goal literature has progressed accordingly. However, we do think that there is room for improvement in the area of measurement and that improvement on this front will both strengthen the empirical base of the achievement goal literature and enhance the prospects for successful application of achievement goal findings to actual achievement settings.

### *Examining the Structure and Predictive Utility of the AGQ-R*

The empirical component of the present research involved testing whether an achievement goal measure that takes into account the aforementioned measurement problems can exhibit good structural validity and have good predictive utility. Accordingly, we examined the factor structure of the AGQ-R using standard confirmatory factor analytic techniques and compared the proposed four-factor model to a series of alternative three- and two-factor models (see Conroy et al., 2003; Elliot & McGregor, 2001). We also examined whether the hypothesized four-factor model was influenced by response bias, which could produce results that appear supportive but in actuality represent artifacts (see Podsakoff, MacKenzie, Lee, & Podsakoff, 2003), and whether inclusion of the same prefixes across goal scales distorted the factor structure of the goals. Most important, we moved beyond the analysis of factor structure to an analysis of dimensional structure, testing for the first time the hypothesis that the four goals represent combinations of two underlying dimensions. This hypothesized model was also compared to alternative dimensional models. Our

Table 1  
*Items for the Achievement Goal Questionnaire-Revised (AGQ-R) Paired With the Original Achievement Goal Questionnaire (AGQ) Items*

Item	Item content
Mastery-approach goal items	
1	My aim is to completely master the material presented in this class. (original Item 9: I desire to completely master the material presented in this class.)
7	I am striving to understand the content of this course as thoroughly as possible. (original Item 8: It is important for me to understand the content of this course as thoroughly as possible.)
3	My goal is to learn as much as possible. (original Item 7: I want to learn as much as possible from this class.)
Mastery-avoidance goal items	
5	My aim is to avoid learning less than I possibly could. (original Item 4: I worry that I may not learn all that I possibly could in this class.)
11	I am striving to avoid an incomplete understanding of the course material. (original Item 5: Sometimes I'm afraid that I may not understand the content of this class as thoroughly as I'd like).
9	My goal is to avoid learning less than it is possible to learn. (original Item 6: I am often concerned that I may not learn all that there is to learn in this class.)
Performance-approach goal items	
4	My aim is to perform well relative to other students. (original Item 3: My goal in this class is to get a better grade than most of the other students.)
2	I am striving to do well compared to other students. (original Item 2: It is important for me to do well compared to others in this class.)
8	My goal is to perform better than the other students. (original Item 1: It is important for me to do better than other students.)
Performance-avoidance goal items	
12	My aim is to avoid doing worse than other students. (original Item 10: I just want to avoid doing poorly in this class.)
10	I am striving to avoid performing worse than others. (original Item 12: My fear of performing poorly in this class is often what motivates me.)
6	My goal is to avoid performing poorly compared to others. (original Item 11: My goal in this class is to avoid performing poorly.)

*Note.* Original AGQ items are from "A  $2 \times 2$  achievement goal framework," by A. J. Elliot and H. A. McGregor, 2001, *Journal of Personality and Social Psychology*, 80, 501–519. Copyright 2001 by the American Psychological Association.

predictions regarding the structure of the AGQ-R were that the 12 items would represent four separable achievement goals emerging from two independent underlying dimensions and that the hypothesized factorial and dimensional structures would prove superior to all alternatives.

An additional aim of our research was to test the new achievement goal measure within the context of the hierarchical model of approach-avoidance achievement motivation. That is, we examined achievement motives as antecedents of achievement goals and the achievement goals as proximal predictors of achievement-relevant outcomes. With regard to the antecedents of achievement goals, achievement motives—the need for achievement and fear of failure—are general, affectively based orientations toward competence presumed to be rooted in early childhood experience (Birney, Burdick, & Teevan, 1969; McClelland, Atkinson, Clark, & Lowell, 1953). We hypothesized that need for achievement would positively predict mastery-approach and performance-approach goals (Tanaka & Yamauchi, 2001; VandeWalle, 1997; Zusho, Pintrich, & Cortina, 2005). Once negative affect is removed from the mastery-avoidance goal items, need for achievement may positively predict these goals as well, although we offer this hypothesis tentatively. Avoiding errors and mistakes is sometimes a necessary component of successfully accomplishing a task; therefore, individuals high in need for achieve-

ment may be more likely to make strategic use of these mastery-avoidance goals than those low in need for achievement. We hypothesized that fear of failure would positively predict performance-approach goals, mastery-avoidance goals, and performance-avoidance goals (Conroy et al., 2003; Elliot & McGregor, 2001; Tanaka, Murakami, Okuno, & Yamauchi, 2002). With regard to the consequences of achievement goal pursuit, we hypothesized that mastery-approach goals would be positive predictors of intrinsic motivation (Harackiewicz, Barron, Tauer, Carter, & Elliot, 2000; Lee, Sheldon, & Turban, 2003; Rawsthorne & Elliot, 1999), whereas mastery-avoidance goals would be negative predictors of intrinsic motivation (Cury, Elliot, Da Fonseca, & Moller, 2006). We hypothesized that performance-approach goals would be positive predictors of performance (Harackiewicz et al., 1997; Urdan, 2004; Wolters, 2004), whereas performance-avoidance goals would be negative predictors of both intrinsic motivation and performance (Elliot & Church, 1997; Finney, Pieper, & Barron, 2004; Pajares & Valiante, 2001).

## Method

### *Participants and Achievement Context*

A total of 229 (76 male, 150 female, and 3 unspecified) undergraduates enrolled in an introductory-level psychology course at a



northeastern university participated in the study in return for extra course credit. Most participants were sophomores or juniors at the university, with a mean age of 19.41 years for the sample. Participants' ethnicity was as follows: 3.93% African American, 16.16% Asian, 68.56% Caucasian, 4.80% Hispanic, 0% Native American, 4.37% other, and 2.18% unspecified. The class was conducted in lecture format, and evaluation was based on a normative grading structure.

### Procedure

The achievement motive and response bias variables were assessed in take-home questionnaire packets during the first 2 weeks of the semester. Participants' achievement goals for their first exam were assessed in a large group session approximately 1 month after the achievement motive and response bias variables were assessed, and 1 week prior to the exam. Intrinsic motivation was assessed in a large group session near the end of the semester, approximately 3 months after the initial take-home assessments. For all assessments, participants were assured that their responses would remain confidential and would in no way influence their course grade. Exam grades were obtained from the professor at the end of the course; participants' SAT scores were obtained from the university registrar.

### Measures

**Need for achievement.** The Achievement Motive subscale of Jackson's (1974) Personality Research Form was used as an indicator of the need for achievement (sample item: "I enjoy difficult work"). A number of empirical investigations have attested to the construct validity of this measure (see Fineman, 1977). Participants responded with *true* (1) or *false* (0) to the 16 items; after reverse scoring eight negatively worded items, the items were summed to form the need for achievement index. Cronbach's  $\alpha$ , computed using the tetrachoric correlation matrix, is .88.

**Fear of failure.** The five-item short form of Conroy's (2001) Performance Failure Appraisal Inventory was used as an indicator of fear of failure (sample item: "When I am failing, I worry about what others think about me"). Several studies have demonstrated the construct validity of this measure (see Conroy & Elliot, 2004). Participants responded on a scale of 1 (*do not believe at all*) to 5 (*believe 100% of the time*), and responses were averaged to form the fear of failure index (Cronbach's  $\alpha = .76$ ).

**Achievement goals.** A series of pilot studies was conducted prior to the present research in the interest of obtaining an optimal set of  $2 \times 2$  achievement goal items. The 12 items listed in Table 1 were eventually selected because they were thought to accurately represent the four goals of the  $2 \times 2$  model while attending carefully to the potential pitfalls reviewed in the introduction. In the present study, the focus of the measure was on students' achievement goals for their first exam in their introductory-level psychology course. Participants responded on a scale of 1 (*strongly disagree*) to 5 (*strongly agree*), and the items were averaged to form the mastery-approach, performance-approach, mastery-avoidance, and performance-avoidance indexes (see Results section for internal consistencies).

**Intrinsic motivation.** Elliot and Church's (1997) eight-item measure was used to assess participants' intrinsic motivation for

the class (sample item: "I think this class is interesting"). The construct validity of this measure has been documented by Elliot and Church (1997). Participants responded on a scale of 1 (*strongly disagree*) to 7 (*strongly agree*); after reverse scoring two negatively worded items, the items were averaged to form the intrinsic motivation index (Cronbach's  $\alpha = .92$ ).

**Response bias.** The 40 items from Paulhus's (1991) Balanced Inventory of Desirable Responding (BIDR) were used to create several different measures of response bias. The BIDR is comprised of two 20-item subscales: Impression Management (IM) and Self-Deceptive Enhancement (SDE). Participants responded to each item using a scale of 1 (*not true*) to 7 (*very true*). Half of the items for each subscale represent desirable statements (e.g., IM: "I always obey laws, even if I'm unlikely to get caught"; SDE: "I always know why I like things"), and half represent undesirable statements (e.g., IM: "When I was young I sometimes stole things"; SDE: "I have not always been honest with myself"). After reverse scoring the undesirable statements, participants received one point for each extreme (6 or 7) response, and their scores for each subscale were summed to form IM and SDE indexes (see Paulhus, 1991, for information on construct validity). An overall BIDR social desirability index was also created by summing the IM and SDE items. Cronbach's  $\alpha$ s for the IM, SDE, and overall BIDR social desirability indexes, computed using the tetrachoric correlation matrices, are .87, .81, and .87, respectively.

Following Elliot and Thrash (2002), we used the BIDR items to also create self-enhancement bias and self-protection bias indexes. Prior to reverse scoring, we summed (across IM and SDE subscales) the number of participants' extreme (6 or 7) responses to the desirable statements and the number of their extreme (1 or 2) responses to the undesirable statements. The first 20-item measure, self-enhancement response bias, represents a tendency to agree with positive statements about oneself that are uncommon, whereas the second 20-item measure, self-protection response bias, represents a tendency to disagree with negative statements about oneself that are common. Cronbach's  $\alpha$ s for these indexes, computed using the tetrachoric correlation matrices, are .78 and .80, respectively.

In addition to the BIDR, the 33-item Marlowe-Crowne Social Desirability Scale (MCSDS; Crowne & Marlowe, 1960) was also used to assess overall social desirability (e.g., "I'm always willing to admit it when I make a mistake"). Participants responded *true* (1) or *false* (0); after reverse scoring 15 negatively worded items, the items were summed to form the MCSDS index. Cronbach's  $\alpha$ , computed using the tetrachoric correlation matrix, is .86.

**Exam performance.** Participants' exam grades were used as a measure of performance attainment. Grades were based on participants' total score on the exam, and each possible grade was assigned a numerical value from 0 to 10 ( $F = 0$ ,  $D = 1$ ,  $D+ = 2$ ,  $C- = 3$ ,  $C = 4$ ,  $C+ = 5$ ,  $B- = 6$ ,  $B = 7$ ,  $B+ = 8$ ,  $A- = 9$ ,  $A = 10$ ). Using total score rather than exam grade as the indicator of performance yielded results essentially identical to those reported in this article.

**SAT scores.** A total SAT score combining the Verbal and Quantitative subscales was used as an index of ability. SAT scores have been shown to predict performance attainment and have been used accordingly as control variables in prior research (Church, Elliot, & Gable, 2001; Elliot & McGregor, 1999).

## Results

*Factorial Structure of Achievement Goals*

Confirmatory factor analyses (CFAs) were conducted on the achievement goal items using AMOS 5.0 (SPSS; Chicago, IL). The analyses were conducted on covariance matrices, and the solutions were generated on the basis of maximum-likelihood estimation. As recommended by Hoyle and Panter (1995), we used several different indexes to evaluate the fit of the models to the data, including chi-square degree of freedom ratio ( $\chi^2/df$ ), comparative fit index (CFI), incremental fit index (IFI), and root-mean-square error of approximation (RMSEA). The following criteria were used to evaluate the adequacy of model fit:  $\chi^2/df \leq 2.0$  (Hair, Anderson, Tatham, & Black, 1995),  $CFI \geq .90$ ,  $IFI \geq .90$ , and  $RMSEA \leq .08$  (Browne & Cudeck, 1993). When multiple models were compared, the Akaike information criterion and Bayesian information criterion were used as well (the lower those values are, the better the fit is).

*Basic CFAs and internal consistencies.* The first CFA examined the hypothesized model, which designated that the items for each goal load on their respective latent factors. To identify the model, the variance of each latent factor was fixed to 1 (Bollen, 1989). The results from this analysis strongly supported the hypothesized model, as not only were all factor loadings quite high (ranging from .93 to .73), but each fit statistic met the criteria for a good fitting model:  $\chi^2(48, N = 229) = 78.32, p < .01, \chi^2/df = 1.63, CFI = .99, IFI = .99, RMSEA = .053$ . All of the subscales demonstrated high levels of internal consistency: For mastery-approach goals, mastery-avoidance goals, performance-approach goals, and performance-avoidance goals, Cronbach's  $\alpha = .84, .88, .92$ , and  $.94$ , respectively.

*Comparison with alternative models.* Additional CFAs investigated the fit of alternative models and compared the fit of the hypothesized and alternative models (see Elliot & McGregor,

2001). Six alternative models were tested: (a) trichotomous model A, in which the performance-approach and performance-avoidance items load on their respective latent factors, and the mastery-approach and mastery-avoidance items load together on a third latent factor; (b) trichotomous model B, in which the mastery-approach and mastery-avoidance items load on their respective latent factors, and the performance-approach and performance-avoidance items load together on a third latent factor; (c) trichotomous model C, in which the mastery-approach and performance-approach items load on their respective latent factors, and the mastery-avoidance and performance-avoidance items load together on a third latent factor; (d) trichotomous model D, in which the mastery-avoidance and performance-avoidance items load on their respective latent factors and the mastery-approach and performance-approach items load together on a third factor; (e) a mastery-performance model in which the mastery-approach and mastery-avoidance items load together on one latent variable, and the performance-approach and performance-avoidance items load together on another; and (f) an approach-avoidance model in which the mastery-approach and performance-approach items load together on one latent variable, and the mastery-avoidance and performance-avoidance items load together on another.

As may be seen in Table 2, the fit indexes indicate that none of the alternative models provided a good fit to the data, and log-likelihood ratio tests show that the hypothesized model provided a far better fit than any of the alternative models. It is worth noting that although the Pearson product moment correlations between some pairs of goal subscales were rather high (i.e., mastery-approach and mastery-avoidance,  $r = .51$ ; mastery-avoidance and performance-avoidance,  $r = .46$ ; performance-approach and performance-avoidance,  $r = .68$ ), our data clearly show that the goals within these pairings are not equivalent.

*CFAs controlling for response bias.* To examine response bias, we created five data sets from the original data set by

Table 2  
*Comparison With the 2 × 2 Model and Alternative Models*

Variable	Overall fit indices					
	$\chi^2/df$	CFI	IFI	RMSEA	AIC	BIC
Hypothesized model	1.63	.99	.99	.053	138.3	241.3
Trichotomous Model A	5.81	.89	.89	.145	350.2	442.9
Trichotomous Model B	7.45	.85	.85	.168	433.8	526.5
Trichotomous Model C	10.10	.78	.78	.200	568.9	661.6
Trichotomous Model D	8.52	.82	.82	.182	488.6	581.3
Mastery-performance model	10.75	.76	.76	.207	619.6	705.6
Approach-avoidance model	14.89	.66	.66	.247	879.2	925.0
Log-likelihood ratio test (model comparison)						
	<i>df</i>	$\chi^2$	<i>p</i>			
Hypothesized model versus						
Trichotomous Model A	3	217.9	< .001			
Trichotomous Model B	3	301.6	< .001			
Trichotomous Model C	3	436.5	< .001			
Trichotomous Model D	3	356.2	< .001			
Mastery-performance model	5	491.3	< .001			
Approach-avoidance model	5	710.8	< .001			

*Note.* CFI = comparative fit index; IFI = incremental fit index; RMSEA = root-mean-square error of approximation; AIC = Akaike information criterion; BIC = Bayesian information criterion.



residualizing various forms of response bias out of each achievement goal item. These new data sets are: (a) an approach-avoidance residualization data set, in which self-enhancement bias scores were residualized out of each mastery-approach and performance-approach item, whereas self-protection bias scores were residualized out of each mastery-avoidance and performance-avoidance item; (b) an SDE residualization data set, in which SDE scores were residualized out of each achievement goal item; (c) an IM residualization data set, in which IM scores were residualized out of each achievement goal item; (d) an overall BIDR residualization data set, in which overall BIDR scores were residualized out of each achievement goal item; and (e) an MCSDS residualization data set, in which MCSDS scores were residualized out of each achievement goal item.

CFAs using each of these data sets indicated that the data were a good fit to the model (Table 3); all factor loadings were highly significant and nearly identical with those from the original data set. These results suggest that the four-factor structure of achievement goals is not an artifact of response bias.

To address the response bias issue more extensively, we examined whether achievement goals and response bias form a common factor (Beretvas, Meyers, & Leite, 2002). First, we constructed a CFA model with five factors: four achievement goal factors and one response bias factor. The achievement goal piece of the model was identical to the CFA model reported above. The response bias factor represented one of six types of response bias (SDE, IM, self-enhancing response bias, self-protecting response bias, overall BIDR, or MCSDS). When the response bias factor reflected SDE, IM, self-enhancement response bias, or self-protection response bias, the response bias latent factor was comprised of two 10-item parcels, randomly selected from their corresponding measures. When response bias reflected overall BIDR, two randomly selected 20-item parcels were used as indicators of the latent factor; and when response bias reflected MCSDS, two randomly selected 16-item and 17-item parcels were used as indicators of the latent factor. These five-factor models were compared with their corresponding four-factor models, created by collapsing one achievement goal factor and the response bias factor into a single factor. With four achievement goals and six response bias indicators, a total of 24 model comparisons were conducted using the log-likelihood ratio test. In each comparison, the five-factor model was found to fit significantly better than the four-factor model,  $\chi^2_{diff}(4,$

$N = 221) \geq 54.61, ps < .001$ , indicating that each of the achievement goal variables is clearly distinct from response bias.

*Investigation of the effect of prefix similarity.* As stated earlier, our achievement goal measure uses three prefixes for each achievement goal, and these prefixes are the same across the four goals. As such, it is possible to control for extraneous phrasing effects, which may afford a more precise assessment of each achievement goal construct. On the other hand, it is possible that items using the same phrase have substantial correlated errors, which would distort our results.

To address this issue, we examined whether assuming a correlation between error variables using the same prefix would improve model fit. Eighteen error correlations (for each prefix, six pairs of items can be made) were investigated one by one with log-likelihood ratio tests. The critical ratio was set to .01 to prevent the inflation of Type I error. The results indicated that none of the 18 correlations between error variables significantly improved the fit of the model. Moreover, in all models, each factor loading remained highly significant and was nearly identical to that with the original model. Therefore, it is clear that our results are not contaminated by systematic error attributable to the common wording of the items.

In sum, the results reported above provide strong support for the four-factor model of achievement goals. Alternative factor models could not explain the data better than the four-factor model. Neither response bias nor prefix similarity distorted the results. Given the good fit of the four-factor model and the high internal consistency of the achievement goal subscales, participants' responses on the items for each hypothesized factor were averaged to form the four achievement goal indexes. Table 4 reports the descriptive statistics of the main variables in this study.

### *The Dimensional Structure of Achievement Goals*

*The multiple-indicator correlated trait-correlated method (MI CT-CM) model.* Although the preceding analyses showed that the four-factor model is highly robust, they do not address the  $2 \times 2$  structure of achievement goals per se. That is, from a theoretical perspective, the valence of competence (positive for approach or negative for avoidance) should be crossed with the definition of competence (mastery or performance), resulting in four separate factors. CFA alone is silent with regard to this dimensional structure. To test the two-dimensional nature of achievement goals, we applied an MI CT-CM model (Marsh & Hocevar, 1988; see also Eid, Lischetzke, Nussbeck, & Trierweiler, 2003) to the data.

An MI CT-CM model is a model typically applied to the multitrait multimethod (MTMM) matrix (Campbell & Fiske, 1959) and is a simple extension of a CT-CM model (Jöreskog, 1974; Marsh & Grayson, 1995). Although our data are not in MTMM format, this model provides us with a way to confirm the dimensionality of the hypothesized constructs. In this model (Figure 1), which can be regarded as akin to a second-order factor analysis model, both the valence and definition dimensions of competence are expected to have additive effects on an achievement goal factor. The valence dimension consists of an approach factor and an avoidance factor, only one of which is applicable to any given goal factor; likewise, the definition dimension consists of a mastery factor and a performance factor, only one of which is applicable to any given goal factor. For example, the mastery-

Table 3  
*Fit Indices for the  $2 \times 2$  Model After Residualizing Response Bias*

Data set	$\chi^2/df$	CFI	IFI	RMSEA
Original data set	1.63	.99	.99	.053
App-av residualization data set	1.66	.99	.99	.055
SDE residualization data set	1.68	.99	.99	.054
IM residualization data set	1.65	.99	.99	.054
Overall BIDR residualization data set	1.65	.98	.98	.056
Overall MCSDS residualization data set	1.70	.98	.98	.056

*Note.* CFI = comparative fit index; IFI = incremental fit index; RMSEA = root-mean-square error of approximation; app-av = approach-avoidance; SDE = self-deceptive enhancement; IM = impression management; BIDR = Paulhus's (1991) Balanced Inventory of Desirable Responding; MCSDS = Marlowe-Crowne Social Desirability Scale.

approach factor is posited to be explained by both the mastery factor and the approach factor. Although factors within each dimension can correlate with each other, it is assumed that factors across dimensions (e.g., the approach factor and the performance factor) are uncorrelated. Thus, the valence and definition dimensions each contribute independently to the achievement goal factors, which allows the achievement goal factor to be decomposed into valence, definition, and unique residual components. This type of model is sometimes called an *additive model*, because each component additively contributes to the construct.

The model was fit to the data using maximum likelihood estimation (Figure 1). To identify the model, we constrained paths from the same second factors to be equal and fixed the covariance between the mastery and performance factors to 0.<sup>2</sup> The model provided a good fit to the data:  $\chi^2(49, N = 229) = 78.54, p < .01$ ,  $\chi^2/df = 1.60$ , CFI = .99, IFI = .99, RMSEA = .051, and all path coefficients were significant. These results nicely support the  $2 \times 2$  dimensionality of the four achievement goals.

*Other alternative models.* Although the MI CT-CM model provided a good fit to the data, other types of models may be examined with regard to the precise structure of achievement goals. Here we describe some plausible alternative models and compare them with the MI CT-CM model.

The first alternative model is a form of two-level model, an example of which is depicted in Figure 2. In this model, the four achievement goal factors themselves make up a two-factor structure. There are two possible types of two-level structures: (a) a mastery–performance two-level model, in which a mastery factor (consisting of mastery-approach and mastery-avoidance factors) and a performance factor (consisting of performance-approach and performance-avoidance factors) are formed as second-order factors; and (b) an approach–avoidance two-level model, in which an approach factor (consisting of mastery-approach and performance-approach factors) and an avoidance factor (consisting of mastery-avoidance and performance-avoidance factors) are formed as second-order factors. The critical difference between the MI CT-CM model and these two-level models is that the two-level models derive factors out of only one dimension of the  $2 \times 2$  model.

We constrained paths from the same second factors to be equal because freeing those parameter estimates makes the solution very unstable (i.e., showing an improper solution and abnormally high standardized factor loadings). Fitting these models to the data, we obtained a worse fit than the MI CT-CM model (Table 5). More-

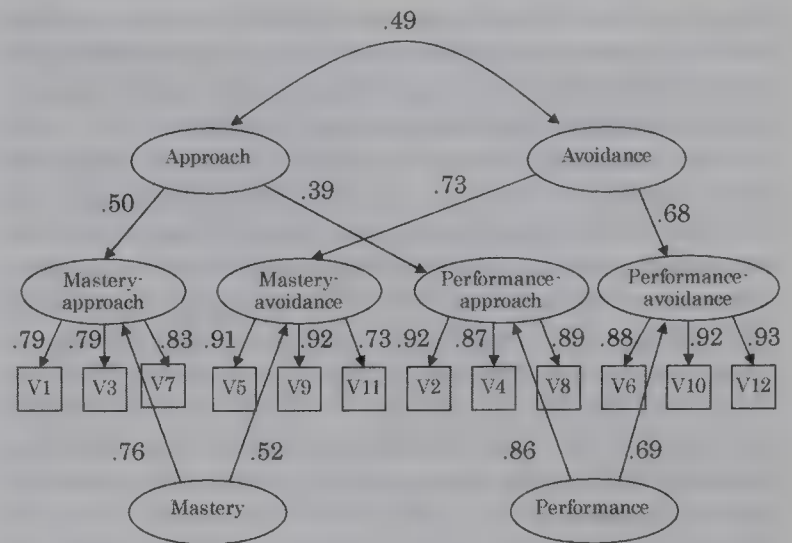


Figure 1. MI CT-CM model of achievement goals. Estimates are standardized. All coefficients are significant ( $p < .01$ ). Error variables are not represented in order to simplify the presentation. V1–V12 represent the individual items of the scale (numbers indicate the order of items in the questionnaire; see Table 1).

over, in the approach–avoidance two-level model, the correlation between the approach and avoidance factors exceeded 1.0.

The second alternative model is a direct product model (Browne, 1989; Wothke & Browne, 1990), which is also typically applied to an MTMM matrix. Whereas a CT-CM model assumes additive (independent) effects of two dimensions, the direct product model assumes only multiplicative effects. That is, the effect of one dimension is posited to completely be a function of the other dimension. For example, the effect of approach factor on mastery-approach factor is assumed to change depending on the strength of the mastery factor. This model is sometimes called a *multiplicative model*, and is compared with an additive model (for further discussion on these two types of model, see Goffin & Jackson, 1992; Hernandez & Gonzalez-Roma, 2002). When there are  $m \times n$  variables comprising the  $m \times n$  dimensional structure, the direct product model expresses the covariance matrix as

$$\Sigma = \mathbf{D}(\Sigma_{D1} \otimes \Sigma_{D2} + \Delta)\mathbf{D},$$

where  $\Sigma_{D1}$  is a  $(m \times m)$  latent variable score correlation matrix of Dimension 1, and  $\Sigma_{D2}$  is a  $(n \times n)$  latent variable score correlation matrix of Dimension 2.  $\mathbf{D}$  is a  $(mn \times mn)$  positive definite diagonal matrix of scale constants.  $\Delta$  is a  $(mn \times mn)$  diagonal matrix of nonnegative unique error components. We set  $m = n = 2$  and applied this covariance structure to the factor correlations of our CFA model. Accordingly, the covariance matrix of our 12 items can be expressed as

$$\Sigma_v = \mathbf{A}\mathbf{D}(\Sigma_{D1} \otimes \Sigma_{D2} + \Delta)\mathbf{D}\mathbf{A}' + \Sigma_E,$$

where  $\mathbf{A}$  is a  $(12 \times 4)$  matrix of factor loadings in the CFA, and  $\Sigma_E$  is a  $(12 \times 12)$  nonnegative diagonal matrix of error variances for the CFA. In our analysis,  $\mathbf{D}$  was set as an identity matrix to identify the model. Maximum likelihood estimates were obtained

<sup>2</sup> This restriction is reasonable because much empirical research has revealed that mastery and performance goals are independent of each other.

Table 4  
Descriptive Statistics

Variable	<i>M</i>	<i>SD</i>	Observed range	Possible range
Need for achievement	9.63	3.47	1.00–16.00	0.00–16.00
Fear of failure	2.99	0.89	1.00–5.00	1.00–5.00
Mastery-approach goals	4.23	0.67	1.67–5.00	1.00–5.00
Mastery-avoidance goals	3.61	0.95	1.00–5.00	1.00–5.00
Performance-approach goals	4.05	0.94	1.00–5.00	1.00–5.00
Performance-avoidance goals	3.83	1.06	1.00–5.00	1.00–5.00
Intrinsic motivation	5.59	1.09	1.00–7.00	1.00–7.00
Exam performance	6.37	2.00	1.00–10.00	0.00–10.00



using Mx (Neale, Boker, Xie, & Maes, 1999). As may be seen in Table 5, this model evidenced a worse fit to the data than the MI CT-CM model; furthermore, three variance components were negative (i.e., the model yielded an improper solution).

In sum, two types of alternative models (a two-level model and a direct product model) were compared with our proposed MI CT-CM model to further explore the nature of the structure of achievement goals. The results were clear: Neither of these alternative models was superior to the MI CT-CM model, both in terms of fit indexes and acceptability of solutions. In other words, our results indicate that achievement goals are comprised of two independent dimensions. Focusing on one of the dimensions (a two-level model) or on the multiplicative effects of the dimensions (a direct product model) cannot precisely capture the structure of achievement goals.

Testing the Hierarchical Model

We used structural equation modeling (SEM) to test the hierarchical model of approach-avoidance achievement motivation, in which achievement goals are posited to represent intermediate variables between motive dispositions and outcome variables such as performance attainment and intrinsic motivation (Elliot, 2006; Elliot & Church, 1997). We began by establishing the measurement portion of our model within a CFA framework (cf. Anderson and Gerbing, 1988) and then proceeded to the full SEM model.

*Measurement model.* The measurement model consisted of the need for achievement, fear of failure, the four achievement goals, intrinsic motivation, and exam performance. Fear of failure and each achievement goal latent variable were represented by the individual items from their respective scales. The need for achievement latent variable was represented by four parcels randomly selected from the 16 items of the scale. Likewise, the intrinsic motivation latent variable was represented by four parcels randomly selected from the 8 items of the scale. Exam performance was an observed variable; SAT scores were residualized out of the exam performance variable to control for the influence of ability.<sup>3</sup> In this model, the loadings between the indicators and the latent factors were freely estimated, and all exogenous variables were free to correlate with each other. The model was a good fit to the data:  $\chi^2(272, N = 229) = 442.16, p < .01, \chi^2/df = 1.62, CFI =$

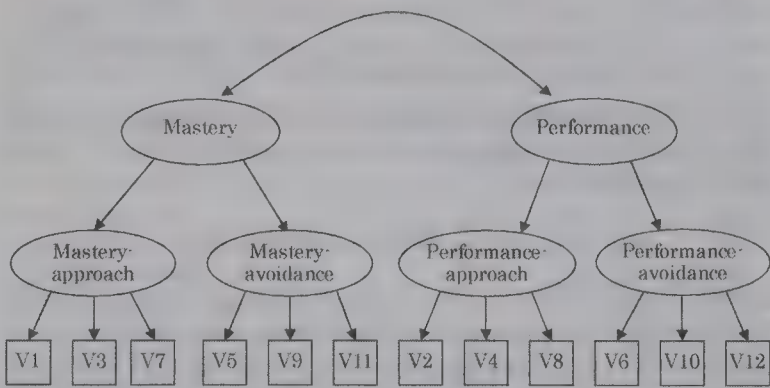


Figure 2. An example of a two-level model of achievement goals. Error variables are not represented in order to simplify the presentation. V1–V12 represent the individual items of the scale (numbers indicate the order of items in the questionnaire; see Table 1).

Table 5  
Fit Indices of CT-CM Model and Other Alternative Models

Model	$\chi^2/df$	CFI	IFI	RMSEA	AIC	BIC
MI CT-CM model	1.60	.99	.99	.051	136.5	236.1
Mastery–performance two-level model	3.18	.95	.95	.098	216.2	308.9
Approach–avoidance two-level model	4.60	.91	.91	.126	290.4	383.1
Direct product model	3.06	—	—	.095	211.2	300.4

Note. CFI = comparative fit index; IFI = incremental fit index; RMSEA = root-mean-square error of approximation; AIC = Akaike information criterion; BIC = Bayesian information criterion; MI CT-CM = multiple-indicator correlated trait–correlated method. Dashes indicate that data were not obtained.

.95, IFI = .95, RMSEA = .052, and all factor loadings were highly significant ( $p < .001$ ).

*Full SEM model.* In the full SEM model, we posited that the achievement motives would lead to achievement goals that, in turn, would directly predict the outcome variables. Specifically, the need for achievement was hypothesized to positively predict mastery-approach goals, performance-approach goals, and, tentatively, mastery-avoidance goals; fear of failure was hypothesized to positively predict mastery-avoidance goals, performance-avoidance goals, and performance-approach goals. Mastery-approach goals were hypothesized to be positive predictors of intrinsic motivation, whereas mastery-avoidance goals were posited to be negative predictors of intrinsic motivation. Performance-approach goals were hypothesized to be positive predictors of exam performance, whereas performance-avoidance goals were posited to be negative predictors of both intrinsic motivation and exam performance. Given the similar wording of the performance-approach and performance-avoidance goal items and the mastery-approach and mastery-avoidance goal items, we allowed for correlated errors between these two latent variables (Elliot & Church, 1997).

In an initial test of the model, all hypothesized paths were significant with the single exception of that between mastery-avoidance goals and intrinsic motivation. The model was thus tested in the final analysis with this path removed. The model provided a good fit to the data:  $\chi^2(288, N = 229) = 550.73, p < .01, \chi^2/df = 1.91, CFI = .92, IFI = .92, RMSEA = .063$ , and all factor loadings were highly significant ( $p < .001$ ). Regarding the antecedents of achievement goals, the need for achievement was a positive predictor of mastery-approach goals ( $\beta = .34, p < .01$ ), performance-approach goals ( $\beta = .22, p < .01$ ), and mastery-avoidance goals ( $\beta = .21, p < .01$ ), whereas fear of failure was a positive predictor of performance-avoidance goals ( $\beta = .31, p < .01$ ), mastery-avoidance goals ( $\beta = .15, p < .05$ ), and performance-approach goals ( $\beta = .24, p < .01$ ). Regarding the consequences of achievement goals, mastery-approach goals were a positive predictor of intrinsic motivation ( $\beta = .28, p < .01$ ), performance-approach goals were a positive predictor of exam

<sup>3</sup> Given that there are some missing values for the SAT scores, the full information maximum likelihood method was used to avoid loss of information due to the missing data (Arbuckle, 1996; Schafer & Graham, 2002).

performance ( $\beta = .46, p < .01$ ), and performance-avoidance goals were a negative predictor of both intrinsic motivation ( $\beta = -.15, p < .05$ ) and exam performance ( $\beta = -.48, p < .01$ ). See Figure 3 for a pictorial summary of these findings.

In sum, SEM analyses yielded data that strongly supported the hierarchical model of approach-avoidance achievement motivation. Both the measurement model and the full SEM model provided a good fit to the data, and the path results in the full SEM model were nearly perfectly consistent with predictions.

### *Comparing the Present Data and the Elliot and McGregor (2001) Data*

Finally, we thought it would be informative to compare the data from the present study with that from the original research with the AGQ (see Elliot & McGregor, 2001). First, we considered the achievement goal reliabilities and the intercorrelations among the achievement goal variables (see Table 6). The internal consistencies for all four goals were strong (greater than .80) in both the present research and the Elliot and McGregor (2001) research. The one noticeable difference was that the  $\alpha$  for performance-avoidance goals was .83 with the AGQ, whereas the  $\alpha$  was .94 for performance-avoidance goals with the AGQ-R. This enhanced reliability is likely due to the explicit normative focus included in all three performance-avoidance goal items in the AGQ-R. The achievement goal intercorrelations were the same in the present research and in the Elliot and McGregor (2001) research, except that mastery-approach and performance-approach goals were significantly positively correlated in the AGQ-R, whereas in the AGQ they were not. This positive correlation is in line with predictions because these two goals share a competence dimension (both are approach goals).

Second, we considered achievement motives as simultaneous predictors of achievement goal adoption (see Table 7). The pre-

dictive patterns were the same in the present research and in the Elliot and McGregor (2001) research, with one exception: Need for achievement was a significant positive predictor of AGQ-R mastery-avoidance goals, but not AGQ mastery-avoidance goals. This positive relation is in line with our predictions regarding mastery-avoidance goals as strategic tools that are sometimes put to effective use by highly achievement-oriented individuals.

Third, we considered the achievement goals as simultaneous predictors of exam performance (see Table 8; intrinsic motivation was not examined in the Elliot & McGregor [2001] research). The predictive patterns were the same in the present research and in the Elliot and McGregor (2001) research.

In sum, comparison of the present data and the Elliot and McGregor (2001) data highlights the value of the AGQ-R measure. This new measure yielded results that replicated all of the important findings from the original research; when the results from the present and prior research diverged, it was the present results that were more in line with our theoretical model. Thus, the AGQ-R appears to illustrate that it is indeed possible to address the measurement problems highlighted at the beginning of this article without sacrificing structural validity and predictive utility.

### Discussion

Sustained progress within a psychological literature requires clear operationalization of the focal constructs in the literature that is consistent with the way that the constructs are conceptualized. In the present research, we have identified several problems with existing measures in the achievement goal literature; some of them seem relatively minor (e.g., some performance-based goal items focus on extreme groups), but others seem to be of considerable importance (e.g., some achievement goal items may not actually focus on goals per se). These problems tend to go unnoticed by researchers or, when noticed, tend to simply be acknowledged in

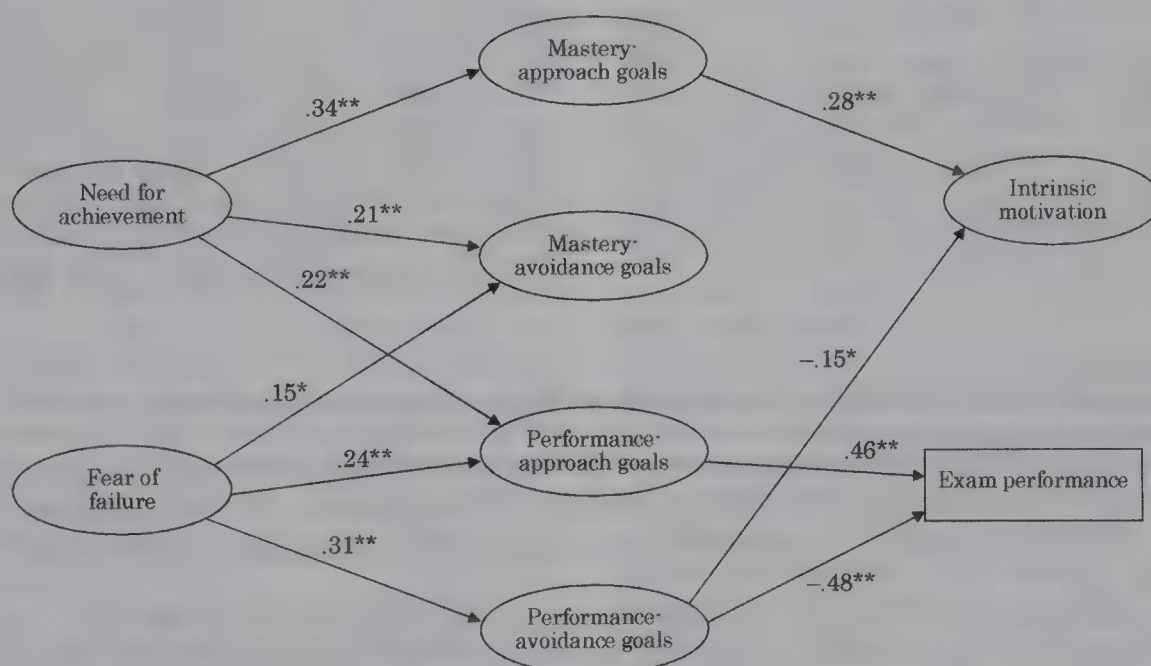


Figure 3. Hierarchical model of achievement goals. Estimates are standardized. Indicator variables, error variables, and correlations between error variables are not represented in order to simplify the presentation. \* $p < .05$ . \*\* $p < .01$ .



Table 6  
*Comparison of Present Results with Elliot and McGregor (2001): Achievement Goal Reliabilities and Intercorrelations*

Goal	1	2	3	4
1. Mastery-approach goals	.84/.88			
2. Mastery-avoidance goals	.51**/.36**	.88/.87		
3. Performance-approach goals	.16*/.07	.15*/.13**	.92/.94	
4. Performance-avoidance goals	.13/-.06	.46**/.29**	.68**/.27**	.94/.83

*Note.* Results from the present study are presented before the slash; results from Elliot and McGregor (2001) are presented after the slash. Results from Elliot and McGregor are aggregated across three studies and examined via meta-analysis (Rosenthal & Rosnow, 1991). Values in the diagonal represent Cronbach's alphas; values in the remainder of the table are Pearson product moment correlation coefficients.  
\**p* < .05. \*\**p* < .01.

the discussion section of empirical articles or noted as an issue to attend to in future research. Our work herein was designed to make these problems salient and to illustrate them by critiquing a commonly used achievement goal measure—the AGQ (Elliot & McGregor, 2001). In addition, our work was designed to attend to these problems in a direct and concrete way by revising the AGQ accordingly. The resulting measure—the AGQ-R—was then put to the test to examine both its structural validity and its predictive utility.

With regard to structural validity, the measure fared extremely well when subjected to a highly rigorous set of analyses. The hypothesized four-factor structure of the achievement goal items was confirmed, and this four-factor structure was found to be a better fit to the data than a series of alternative models with three- and two-factor structures. Additional analyses demonstrated that the four-factor structure was not a mere artifact of self-enhancement, self-protection, self-deception, impression management, or general social desirability. Each of the four achievement goal factors had a high degree of internal consistency. We not only confirmed the four-factor structure in our analyses but also confirmed the two-dimensional structure posited by the 2 (definition) × 2 (valence) achievement goal model. This two-dimensional model also was shown to fit the data better than two types of alternative models.

Table 7  
*Comparison of Present Results with Elliot and McGregor (2001): Achievement Motives as Predictors of Achievement Goals*

Achievement motive	Mastery-approach goals	Mastery-avoidance goals	Performance-approach goals	Performance-avoidance goals
Need for achievement	.28**/.17*	.14*/.00	.27**/.38**	.12/-.02
Fear of failure	.07/-.12	.15*/.28**	.18*/.32**	.26**/.31**

*Note.* Results from the present study are presented before the slash; results from Elliot and McGregor (2001) are presented after the slash. All values represent standardized simultaneous regression coefficients estimated from Pearson product moment correlation coefficients.  
\**p* < .05. \*\**p* < .01.

Table 8  
*Comparison of Present Results With Elliot and McGregor (2001): Achievement Goals as Predictors of Exam Performance*

Achievement goal	Exam performance
Mastery-approach goals	-.06/.10
Mastery-avoidance goals	-.07/-.05
Performance-approach goals	.36**/.20*
Performance-avoidance goals	-.33**/-.31**

*Note.* Results from the present study are presented before the slash; results from Elliot and McGregor (2001) are presented after the slash. All values represent standardized simultaneous regression coefficients estimated from Pearson product moment correlation coefficients.  
\**p* < .05. \*\**p* < .01.

With regard to predictive utility, results from SEM analyses were strongly supportive. The measurement model was a good fit to the data, as was the path model in which achievement motives were posited as predictors of achievement goals and the goals, in turn, were posited as direct predictors of achievement-relevant outcomes. Mastery-approach and performance-avoidance goals were shown to emerge from a single antecedent, the need for achievement and fear of failure, respectively, whereas performance-approach and mastery-avoidance goals were shown to emerge from both of these achievement motives. Mastery-approach goals were positive predictors and performance-avoidance goals were negative predictors of intrinsic motivation; performance-approach goals were positive predictors and performance-avoidance goals were negative predictors of exam performance. These results were nearly perfectly in accord with our hypotheses; the only hypothesized relation that was not supported was that between mastery-avoidance goals and intrinsic motivation (to be discussed in more detail below). In short, analyses focused on both structural validity and predictive utility yielded strong support for the AGQ-R; the measure appears to be empirically as well as conceptually sound. In addition, comparison of the AGQ-R data from the present research with the AGQ data from prior research casts the AGQ-R in a very positive light.

An important feature of the present research was our examination of the dimensional as well as factorial structure of achievement goals. Our dimensional findings are noteworthy because they are the first empirical evidence indicating that each of the four goals of the 2 × 2 model indeed represents a combination of two underlying competence dimensions. Thus, although each of the

goals is a unique combination of dimensions and therefore is distinct, some goals share a dimension and therefore are conceptually related, whereas others do not share a dimension and therefore are conceptually unrelated. The two goals that are composed of completely different dimensions—mastery-approach and performance-avoidance goals—exhibited starkly different empirical profiles in our research. This is true with regard to both antecedents (they emerged from completely different achievement motives) and consequences (they had completely different effects on intrinsic motivation and exam performance).

It is interesting that even among achievement goals that share a dimension there is variability in the strength of relation between them. For example, in our research, performance-approach and performance-avoidance goals were strongly correlated ( $r = .68$ ), whereas mastery-approach and performance-approach goals were not ( $r = .16$ ). One observation on the observed pattern of correlations is that goals sharing a common definition dimension appear to be more closely related than goals sharing a common valence dimension. Another observation is that the strongest link between any two goals appears to be that between performance-approach and performance-avoidance goals. This link is of particular interest because it has led some to question whether these two forms of regulation can truly be separated at the phenomenological level (Roeser, 2004; Urdan & Mestas, 2006) and has led others to fear that performance-approach goals quickly transform into performance-avoidance goals once failure or difficulty is encountered (Brophy, 2005; Midgley et al., 2001). It is possible, however, that the adoption of performance-approach goals evokes perceptual-cognitive processes that are not only functionally and experientially separate from avoidance concerns (Cacioppo, Gardner, & Berntson, 1997; Elliot, 2006) but also bias information processing toward information and interpretation that sustains an approach form of regulation, even in the face of failure (Elliot & Harackiewicz, 1996; Kunda, 1990). Research is clearly needed to directly address this important issue. In addition, a fascinating avenue for research would be to explore moderators of the relation between performance-approach and performance-avoidance goal adoption. We suspect that the joint adoption of these goals is most prevalent in highly evaluative achievement environments and among people with high fear of failure or low competence perceptions.

Mastery-avoidance goals are the most recent addition to the achievement goal literature and are the least researched and least understood of the four goals in the  $2 \times 2$  model. Our antecedent results shed some light on the nature of these goals, as they were found to emerge from both the need for achievement and fear of failure. Prior research has linked mastery-avoidance goals to fear of failure alone (Conroy & Elliot, 2004; Elliot & McGregor, 2001); it took a cleansing of negative affect from the original mastery-avoidance items for the conceptually sensible link between the need for achievement and these mastery-based goals to appear. This, of course, nicely illustrates the empirical value of attending to the conceptual problems plaguing the AGQ and other existing achievement goal measures.

Although clarity was obtained regarding the antecedents of mastery-avoidance goals, clarity remained elusive regarding the consequences of these goals, as they were not negative predictors of intrinsic motivation as anticipated (see Cury et al., 2006). As a combination of the most positive component of achievement goals (mastery) and the most negative (avoidance), mastery-avoidance

goals represent a puzzling motivational hybrid, and it simply is not clear how these two seemingly discordant components operate together in the process of goal regulation. Perhaps in some achievement contexts or for some persons the mastery component of the goal predominates, leading to relatively positive consequences, whereas in other contexts or for other persons the avoidance component predominates, leading to relatively negative consequences. And perhaps most often, the positive and negative components of the goal cancel each other out, leading to neither positive nor negative consequences (as found in the present work). The conceptual and empirical complexities inherent in mastery-avoidance goals can be frustrating at this early stage of study, and it appears that the adoption and pursuit of a mastery-approach goal (with accompanying persistence and task absorption) will be needed for researchers exploring this currently opaque form of regulation.

Although the revised mastery-avoidance goal scale yielded results different from those obtained with the original scale, it is important to note that the revised scales for the other three goals yielded results fully in accord with those from the original scales (and other measures focused on these goals as well). This is reassuring on two fronts. First, and most specifically, it highlights the utility of achievement goals per se as predictors of important achievement-relevant outcomes. Second, and more generally, it indicates that the problems identified herein with existing measures are not so severe as to produce a completely spurious empirical corpus. Indeed, it is difficult to imagine the achievement goal approach rising to its lofty status in the field unless the existing measures captured, to a great extent, systematic and conceptually relevant variance. However, we do think that the problems that we have identified muddy the waters by adding unsystematic and conceptually irrelevant variance to the assessment process and that this muddying of the waters has kept the achievement goal approach from developing to its full potential. It is our hope that additional clarity and precision on the measurement front will translate into additional clarity and precision on the empirical front and will help the achievement goal literature move forward.

We view our revised achievement goal measure as an improvement on the AGQ, but we in no way view the AGQ-R as some sort of final or definitive assessment tool. On the contrary, in the achievement goal literature and in psychological literatures more generally, we view measurement development as part and parcel of theory development, such that with advances or refinements in theory comes the need for revised or new measures. For example, Elliot (1999) stated that mastery-based goals may be differentiated in terms of whether a task-based or intrapersonal standard is used in competence evaluation. The AGQ-R is somewhat ambiguous on this issue, although the most straightforward reading of the items suggests a task-based focus (see Van Yperen, 2006, for mastery-based goals that use an intrapersonal focus). Embracing this  $3 \times 2$  achievement goal framework will necessitate the development of a new measure that explicitly differentiates task-based mastery goals from intrapersonal mastery goals. We think the time has come to move in this direction, as we think it highly likely that each of the six goals in the  $3 \times 2$  model will indeed demonstrate factorial separability and discriminant predictive utility.

It should also be highlighted that there is room for disagreement as to how achievement goals should be conceptualized, particu-



larly with regard to the performance-mastery distinction, and this clearly has implications for achievement goal measurement. One ongoing issue is whether performance-based goal measures should include a demonstration component as well as a normative component (Elliot, 2006; Urdan & Mestas, 2006). The AGQ-R focuses exclusively on the normative component, leaving the demonstration component as an optional feature that may appear in performance-based goal complexes (see Elliot & Thrash, 2001) but need not be present in all performance-based goal pursuit. Another ongoing issue, albeit one that has received less attention, is whether the focal point of performance and mastery goals should be the same or different. The AGQ-R follows convention in having performance-based goals focus on performing and having mastery-based goals focus on learning (see Dweck, 1986), but it would be possible (although, we suspect, not desirable for most) to devise items that eliminated this distinction altogether. Finally, we should acknowledge that our research focused on college undergraduates in a classroom context; the generalizability of our results to other ages and contexts remains an open question.

One reason for explicating the measurement problems present in the achievement goal literature is to highlight the need for improved measures, the use of which, it is hoped, will improve the quality of the empirical literature and the effectiveness of applied endeavors. However, we close by noting another, broader reason for striving for improved achievement goal assessment. Achievement goals are but one of the many constructs necessary to fully account for behavior in achievement settings. Other constructs such as motives, values, emotions, and competence perceptions, to name a few, are also important to include in models of achievement motivation (for a review of constructs, see Anderman & Wolters, 2006; Elliot & Dweck, 2005). Existing measures of achievement goals often include content that more readily belongs in measures of these other constructs. As the achievement goal literature matures it will undoubtedly move more and more in the direction of linking goals to these other constructs in integrative fashion. A conceptually clean achievement goal measure would seem a prerequisite for this integration process to transpire smoothly.

## References

- Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*, 261–271.
- Anderman, E. M., & Wolters, C. (2006). Goals, values, and affect. In P. Alexander & P. Winne (Eds.), *Handbook of educational psychology* (2nd ed., pp. 369–390). Mahwah, NJ: Erlbaum.
- Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103*, 411–423.
- Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243–277). Mahwah, NJ: Erlbaum.
- Barron, K. E., & Harackiewicz, J. M. (2001). Achievement goals and optimal motivation: Testing multiple goal models. *Journal of Personality and Social Psychology, 80*, 706–722.
- Beretvas, N. S., Meyers, J. L., & Leite, W. L. (2002). A reliability generalization study of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement, 4*, 570–589.
- Birney, R. C., Burdick, H., & Teevan, R. C. (1969). *Fear of failure*. New York: Nostrand.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Bouffard, T., Boisvert, J., Vezeau, C., & Larouche, C. (1995). The impact of goal orientation on self-regulation and performance among college students. *British Journal of Educational Psychology, 65*, 317–329.
- Brophy, J. (2005). Goal theorists should move on from performance goals. *Educational Psychologist, 40*, 167–176.
- Browne, M. W. (1989). Relationships between an additive model and a multiplicative model for multitrait-multimethod matrices. In R. Coppi & S. Bolasco (Eds.), *Multitrait data analysis* (pp. 507–520). New York: Elsevier.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverley Hills, CA: Sage.
- Button, S. B., Mathieu, J. E., & Zajac, D. M. (1996). Goal orientation in organizational research: A conceptual and empirical foundation? *Organizational Behavior and Human Decision Processes, 67*, 26–48.
- Cacioppo, J. T., Gardner, W. L., & Berntson, G. G. (1997). Beyond bipolar conceptualization and measures: The case of attitudes and evaluative space. *Personality and Social Psychology Review, 1*, 3–25.
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.
- Church, M. A., Elliot, A. J., & Gable, S. L. (2001). Perceptions of classroom environment, achievement goals, and achievement outcomes. *Journal of Educational Psychology, 93*, 43–54.
- Conroy, D. E. (2001). Progress in the development of a multidimensional measure of fear of failure: The Performance Failure Appraisal Inventory (PFAI). *Anxiety, Stress, and Coping: An International Journal, 14*, 431–452.
- Conroy, D. E., & Elliot, A. J. (2004). Fear of failure and achievement goals in sport: Addressing the issue of the chicken and the egg. *Anxiety, Stress, and Coping: An International Journal, 17*, 271–285.
- Conroy, D. E., Elliot, A. J., & Hofer, S. M. (2003). A 2 × 2 achievement goals questionnaire for sport. *Journal of Sport and Exercise Psychology, 25*, 456–476.
- Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24*, 349–354.
- Cury, F., Elliot, A. J., Da Fonseca, D., & Moller, A. (2006). The social-cognitive model of achievement motivation and the 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology, 90*, 666–679.
- Custers, R., & Aarts, H. (2005). Positive affect as implicit motivator: On the nonconscious operation of behavioral goals. *Journal of Personality and Social Psychology, 89*, 129–142.
- Duda, J. L. (2005). Motivation in sport: The relevance of competence and achievement goals. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 318–335). New York: Guilford Press.
- Duda, J. L., Chi, L., Newton, M. L., Walling, M. D., & Catley, D. (1995). Task and ego orientation and intrinsic motivation in sport. *International Journal of Sport Psychology, 26*, 40–63.
- Duda, J. L., & Nicholls, J. G. (1992). Dimensions of achievement motivation in schoolwork and sport. *Journal of Educational Psychology, 84*, 290–299.
- Dweck, C. S. (1986). Motivational process affects learning. *American Psychologist, 41*, 1010–1018.
- Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. Philadelphia: Psychology Press.
- Dweck, C. S., & Elliott, E. (1983). Achievement motivation. In P. H. Mussen (Series Ed.) & E. M. Hetherington (Vol. Ed.), *Handbook of child psychology: Vol. 4. Socialization, personality, and development* (4th ed., pp. 643–691). New York: Wiley.
- Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003).

- Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple-indicator CT-C(M-1) model. *Psychological Methods*, 8, 38–60.
- Elliot, A. J. (1997). Integrating “classic” and “contemporary” approaches to achievement motivation: A hierarchical model of approach and avoidance achievement motivation. In P. Pintrich & M. Maehr (Eds.), *Advances in motivation and achievement* (Vol. 10, pp. 143–179). Greenwich, CT: JAI Press.
- Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist*, 34, 149–169.
- Elliot, A. J. (2005). A conceptual history of the achievement goal construct. In A. Elliot & C. Dweck (Eds.), *Handbook of competence and motivation* (pp. 52–72). New York: Guilford Press.
- Elliot, A. J. (2006). The hierarchical model of approach–avoidance motivation. *Motivation and Emotion*, 30, 111–116.
- Elliot, A. J., & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology*, 72, 218–232.
- Elliot, A. J., & Dweck, C. S. (2005). Competence and motivation. In A. Elliot & C. Dweck (Eds.), *Handbook of competence and motivation* (pp. 3–12). New York: Guilford Press.
- Elliot, A. J., & Fryer, J. W. (2008). The goal concept in psychology. In J. Shah & W. Gardner (Eds.), *Handbook of motivational science* (pp. 235–250). New York: Guilford Press.
- Elliot, A. J., & Harackiewicz, J. M. (1996). Approach and avoidance achievement goals and intrinsic motivation: A mediational analysis. *Journal of Personality and Social Psychology*, 70, 461–475.
- Elliot, A. J., & McGregor, H. A. (1999). Test anxiety and the hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology*, 76, 628–644.
- Elliot, A. J., & McGregor, H. A. (2001). A 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology*, 80, 501–519.
- Elliot, A. J., & Thrash, T. M. (2001). Achievement goals and the hierarchical model of achievement motivation. *Educational Psychology Review*, 12, 139–156.
- Elliot, A. J., & Thrash, T. M. (2002). Approach–avoidance motivation in personality: Approach and avoidance temperaments and goals. *Journal of Personality and Social Psychology*, 82, 804–818.
- Fineman, S. (1977). The achievement motivation construct and its measurement: Where are we now? *British Journal of Psychology*, 68, 1–22.
- Finney, S. J., Pieper, S. L., & Barron, K. E. (2004). Examining the psychometric properties of the achievement goal questionnaire in a general academic context. *Educational and Psychological Measurement*, 64, 365–382.
- Goffin, R. D., & Jackson, D. (1992). Analysis of multitrait–multirater performance appraisal data: Composite direct product method versus confirmatory factor analysis. *Multivariate Behavioral Research*, 27, 363–385.
- Greene, B. A., & Miller, R. B. (1996). Influences on achievement: Goals, perceived ability, and cognitive engagement. *Contemporary Educational Psychology*, 21, 181–192.
- Hair, J. F. Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis with readings* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
- Harackiewicz, J. M., Barron, K. E., Elliot, A. J., Carter, S. M., & Lehto, A. T. (1997). Predictors and consequences of achievement goals in the college classroom: Maintaining interest and making the grade. *Journal of Personality and Social Psychology*, 73, 1284–1295.
- Harackiewicz, J. M., Barron, K. E., Tauer, J. M., Carter, S. M., & Elliot, A. J. (2000). Short-term and long-term consequences of achievement goals: Predicting interest and performance over time. *Journal of Educational Psychology*, 92, 316–330.
- Hernandez, A., & Gonzalez-Roma, V. (2002). Analysis of multitrait–multioccasion data: Additive versus multiplicative models. *Multivariate Behavioral Research*, 37, 59–87.
- Hoyle, R., & Panter, A. (1995). Writing about structural equation models. In R. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 100–119). Thousand Oaks, CA: Sage.
- Jöreskog, K. G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2, pp. 1–56). San Francisco: Freeman.
- Jackson, D. (1974). *Personality Research Form manual*. Goshen, NY: McGraw-Hill.
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108, 480–498.
- Lee, F. K., Sheldon, K. M., & Turban, D. B. (2003). Personality and the goal striving process: The influence of achievement goal patterns, goal level, and mental focus on performance and enjoyment. *Journal of Applied Psychology*, 88, 256–265.
- Maehr, M. L. (1983). On doing well in science: Why Johnny no longer excels; why Sarah never did. In S. Paris, G. Olson, & N. Stephenson (Eds.), *Learning and motivation in the classroom* (pp. 178–210). Hillsdale, NJ: LEA.
- Maehr, M. L. (1989). Thoughts about motivation. In C. Ames & R. Ames (Eds.), *Research on motivation in education* (Vol. 3, pp. 299–315). New York: Academic Press.
- Maehr, M. L., & Nicholls, J. G. (1980). Culture and achievement motivation: A second look. In N. Warren (Ed.), *Studies in cross-cultural psychology* (Vol. 3, pp. 221–267). New York: Academic Press.
- Marsh, H. W., & Grayson, D. (1995). Latent variable models of multitrait–multimethod data. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 177–198). Thousand Oaks, CA: Sage.
- Marsh, H. W., & Hocevar, D. (1988). A new, more powerful approach to multitrait–multimethod analyses: Application of second order confirmatory factor analysis. *Journal of Applied Psychology*, 73, 107–117.
- McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. (1953). *The achievement motive*. New York: Appleton-Century-Crofts.
- Meece, J., Anderman, E. M., & Anderman, L. H. (2006). Classroom goal structure, student motivation, and academic achievement. *Annual Review of Psychology*, 57, 505–528.
- Meece, J. L., Blumenfeld, P. C., & Hoyle, R. H. (1988). Students’ goal orientations and cognitive engagement in classroom activities. *Journal of Educational Psychology*, 80, 514–523.
- Middleton, M. J., & Midgley, C. (1997). Avoiding the demonstration of lack of ability: An underexplored aspect of goal theory. *Journal of Educational Psychology*, 89, 710–718.
- Midgley, C., Kaplan, A., & Middleton, M. (2001). Performance approach goals: Good for what, for whom, under what circumstances, and at what cost? *Journal of Educational Psychology*, 93, 77–86.
- Midgley, C., Maehr, M. L., Huda, L. Z., Anderman, E., Anderman, L., Freeman, K. E., et al. (2000). *Manual for the Patterns of Adaptive Learning Scales (PALS)*. Ann Arbor: University of Michigan.
- Neale, M. C., Boker, S. M., Xie, G., & Maes, H. H. (1999). *Mx: Statistical modeling* (5th ed.). Richmond: Department of Psychiatry, Medical College of Virginia.
- Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological Review*, 91, 328–346.
- Nicholls, J. G. (1989). *The competitive ethos and democratic education*. Cambridge, MA: Harvard University Press.
- Nicholls, J. G., Cobb, P., Wood, T., Yackel, E., & Patashnick, M. (1990). Assessing students’ theories of success in mathematics: Individual and classroom differences. *Journal for Research in Mathematics Education*, 21, 109–122.



- Nicholls, J. G., Patashnick, M., & Nolen, S. B. (1985). Adolescents' theories of education. *Journal of Educational Psychology*, 77, 683-692.
- Pajares, F., & Valiante, G. (2001). Gender differences in writing motivation and achievement of middle school students: A function of gender orientation? *Contemporary Educational Psychology*, 20, 366-381.
- Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17-59). San Diego, CA: Academic Press.
- Payne, S. C., Youngcourt, S. S., & Beaubien, J. M. (2007). A meta-analytic examination of the goal orientation nomological net. *Journal of Applied Psychology*, 92, 128-150.
- Pintrich, P. R. (2000). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology*, 92, 544-555.
- Pintrich, P. R., & Schunk, D. H. (1996). *Motivation in education: Theory, research, and applications*. Englewood Cliffs, NJ: Prentice Hall.
- Podsakoff, P. M., MacKenzie, S. M., Lee, J., & Podsakoff, N. P. (2003). Common method variance in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88, 879-903.
- Rawsthorne, L. J., & Elliot, A. J. (1999). Achievement goals and intrinsic motivation: A meta-analytic review. *Personality and Social Psychology Review*, 3, 326-344.
- Roberts, G. C., & Treasure, D. C. (1995). Achievement goals, motivational climate and achievement strategies and behaviors in sport. *International Journal of Sport Psychology*, 26, 64-80.
- Roedel, T. D., Schraw, G., & Plake, B. S. (1994). Validation of a measure of learning and performance goal orientations. *Educational and Psychological Measurement*, 54, 1013-1021.
- Roeser, R. W. (2004). Competing schools of thought in achievement goal theory? In M. L. Maehr & P. R. Pintrich (Eds.), *Advances in motivation and achievement: Vol. 13. Motivating students, improving schools: The legacy of Carol Midgley* (pp. 265-299). New York: Elsevier.
- Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw-Hill.
- Ryan, K., Ryan, A. M., Arbuthnot, K., & Samuels, M. (2007). Students' motivation for standardized math exams. *Educational Researcher*, 36, 1-9.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147-177.
- Skaalvik, E. M. (1997). Self-enhancing and self-defeating ego orientation: Relations with task and avoidance orientation, achievement, self-perceptions, and anxiety. *Journal of Educational Psychology*, 89, 71-81.
- Stipek, D., & Gralinski, J. H. (1996). Children's beliefs about intelligence and school performance. *Journal of Educational Psychology*, 88, 397-407.
- Tanaka, A., Murakami, Y., Okuno, T., & Yamauchi, H. (2002). Achievement goals, attitudes toward help seeking, and help-seeking behavior in the classroom. *Learning and Individual Differences*, 13, 23-35.
- Tanaka, A., & Yamauchi, H. (2001). A model for achievement motives, goal orientations, intrinsic interest, and academic achievement. *Psychological Reports*, 88, 123-135.
- Thrash, T. M., & Elliot, A. J. (2001). Delimiting and integrating the goal and motive constructs in achievement motivation. In A. Efklides, J. Kuhl, & R. Sorrentino (Eds.), *Trends and prospects in motivation research* (pp. 3-21). Amsterdam: Kluwer.
- Urdu, T. (2004). Predictors of academic self-handicapping and achievement: Examining achievement goals, classroom goal structures, and culture. *Journal of Educational Psychology*, 96, 251-264.
- Urdu, T., & Mestas, M. (2006). The goals behind performance goals. *Journal of Educational Psychology*, 98, 354-365.
- VandeWalle, D. (1997). Development and validation of a work domain goal orientation instrument. *Educational and Psychological Measurement*, 57, 995-1015.
- Van Yperen, N. W. (2006). A novel approach to assessing achievement goals in the context of the 2 x 2 framework: Identifying distinct profiles of individuals with different dominant achievement goals. *Personality and Social Psychology Bulletin*, 32, 1432-1445.
- Wolters, C. A. (2004). Advancing achievement goals theory: Using goal structures and goal orientations to predict students' motivation, cognition, and achievement. *Journal of Educational Psychology*, 96, 236-250.
- Wothke, W., & Browne, W. W. (1990). The direct product model for the MTMM matrix parameterized as a second order factor analysis model. *Psychometrika*, 55, 255-262.
- Zusho, A., Pintrich, P. R., & Cortina, K. S. (2005). Motives, goals, and adaptive patterns of performance in Asian American and Anglo American students. *Learning and Individual Differences*, 15, 141-158.
- Zweig, D., & Webster, J. (2004). Validation of a multidimensional measure of goal orientation. *Canadian Journal of Behavioural Science*, 36, 232-248.

Received April 24, 2007

Revision received February 5, 2008

Accepted February 26, 2008 ■

# The Relationships Among Students' Future-Oriented Goals and Subgoals, Perceived Task Instrumentality, and Task-Oriented Self-Regulation Strategies in an Academic Environment

Sharon E. Tabachnick  
University of Memphis

Raymond B. Miller  
University of Oklahoma, Norman

George E. Relyea  
University of Memphis

The authors performed path analysis, followed by a bootstrap procedure, to test the predictions of a model explaining the relationships among students' distal future goals (both extrinsic and intrinsic), their adoption of a middle-range subgoal, their perceptions of task instrumentality, and their proximal task-oriented self-regulation strategies. The model was based on R. B. Miller and S. J. Brickman's (2004) conceptualization of future-oriented motivation and self-regulation, which draws primarily from social-cognitive and self-determination theories. Participants were 421 college students who completed a questionnaire that included scales measuring the 5 variables of interest. Data supported the model, suggesting that students' distal future goals (intrinsic future goals in particular) may be related to their middle-range college graduation subgoal, to their perceptions of task instrumentality, and to their adoption of proximal task-oriented self-regulation strategies.

**Keywords:** future and proximal goals, perceived task instrumentality, social-cognitive theory, self-determination theory, intrinsic and extrinsic goals

Miller and Brickman (2004) proposed a model of future-oriented motivation and self-regulation that had the expressed purpose of integrating future and proximal motivation and self-regulation variables. This model drew heavily from social-cognitive theory (Bandura, 1997), self-determination theory (Deci & Ryan, 2000; Ryan & Deci, 2000), personal investment theory (Maehr, 1984), future-time perspective theory (Nuttin, 1985), and the future-oriented extension of achievement motivation theory (Raynor, 1974; Raynor & Entin, 1982). Miller and Brickman's (2004) model, depicted in Figure 1, consists of two major interconnected parts, *future-oriented regulation* and *proximal self-regulation processes*. The full model has been described in detail elsewhere (Miller & Brickman, 2004; Tabachnick, 2005). The backbone of the model is delineated by four major variables (see circled variables in Figure 1) that clearly connect future goals with proximal subgoals, with perceived task instrumentality, and finally, with proximal task-oriented self-regulation. The purpose of the present study was to test a specific portion of Miller and Brickman's (2004) model

to examine the relationships among the four major variables explaining the connection between the hypothesized future and proximal motivation and self-regulation processes. The four variables of interest in the present study were future goals, proximal subgoals, perceived task instrumentality, and task-oriented self-regulation strategies (see Figure 1).

Miller and Brickman (2004) contended that students' distal future goals (e.g., personal growth, contribution to community, personal relationships, etc.) influence the adoption of proximal subgoals in the service of the future goals; that the proximal subgoals lead to perceptions of task instrumentality on the part of students exposed to learning tasks; and that perceived task instrumentality, in turn, leads to proximal task-oriented self-regulation. Although Miller and Brickman (2004) defined personally valued future goals in terms of Deci and Ryan's (2000; also Ryan & Deci, 2000) self-determined future aspirations, they stopped short of differentially modeling future goals in an extrinsic versus intrinsic manner, as Kasser and Ryan (1993, 1996) did.

However, the picture that is beginning to emerge from recent research is that aspiring to attain intrinsic and extrinsic future goals has an important impact on well-being, the quality of task engagement, and achievement. For example, Kasser and Ryan (1993, 1996) have found that aspirations for money and wealth (considered "extrinsic" by self-determination theory) are associated with decreased well-being and mental health in comparison with aspirations for benefitting community and personal growth (considered

---

Sharon E. Tabachnick, Ned R. McWherter Library, University of Memphis; Raymond B. Miller, Department of Educational Psychology, University of Oklahoma, Norman; George E. Relyea, Center for Community Health, University of Memphis.

Correspondence concerning this article should be addressed to Sharon E. Tabachnick, P.O. Box 770126, Memphis, TN 38177-0126. E-mail: sharon.tabachnick@memphis.edu



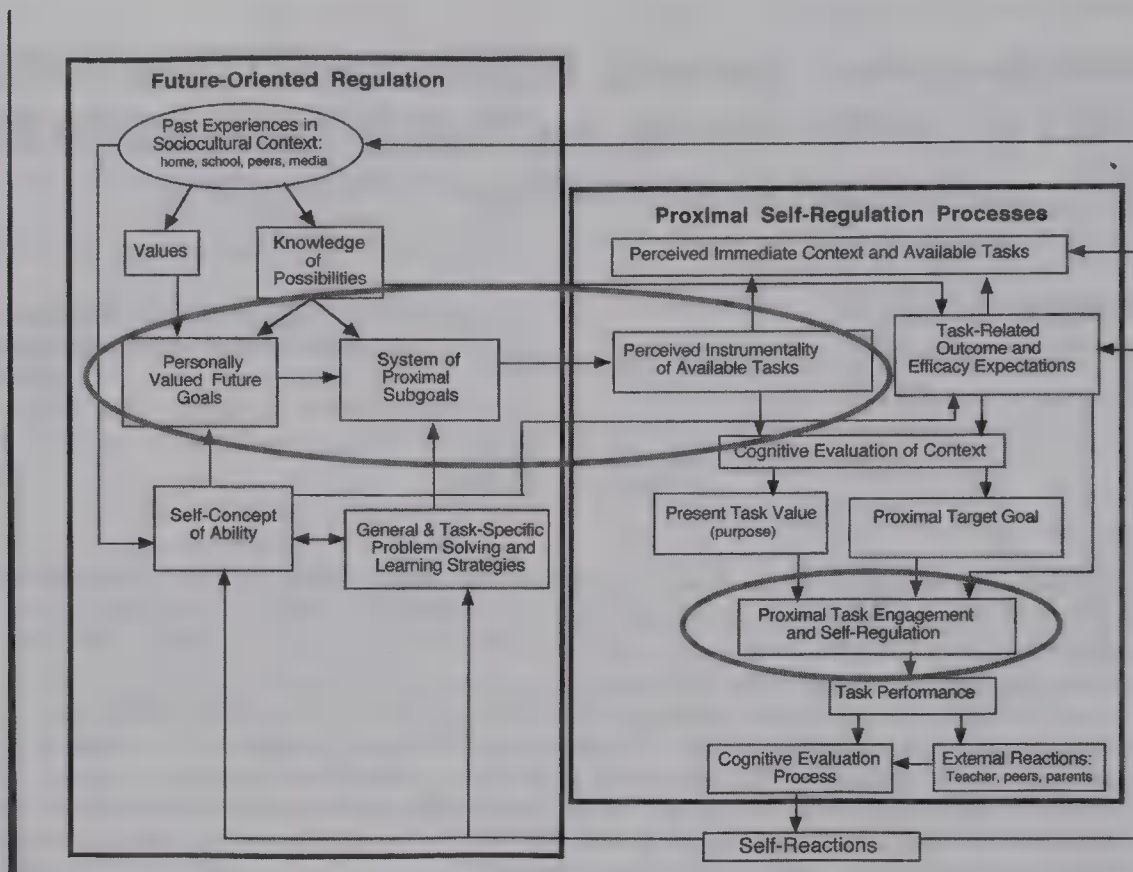


Figure 1. Model of future-oriented motivation and self-regulation, with the variables of interest to the present study circled. From "A model of future-oriented motivation and self-regulation," by R. B. Miller and S. J. Brickman, 2004, *Educational Psychology Review*, 16, p. 13, Figure 1. Copyright 2004 by Springer/Kluwer Academic Publishers. Reproduced and adapted with kind permission of Springer Science and Business Media.

"intrinsic" by self-determination theory).<sup>1</sup> Simons, Dewitte, and Lens (2000, 2004) found, among other things, that students performing tasks because of intrinsic future consequences (e.g., personal growth) were significantly more likely to adopt task or mastery goals and, in addition, had more interest and confidence in their studies, persisted longer, used more deep learning strategies, and received better exam scores than students who performed tasks because of extrinsic future consequences (e.g., high earnings). In a series of studies about experimentally induced goals, Vansteenkiste, Simons, Lens, Sheldon, and Deci (2004) and Vansteenkiste, Simons, Soenens, and Lens (2004) similarly found that framing tasks in terms of intrinsic future goals (e.g., the task is important for personal growth or health in the future) resulted in better test performance, a significantly larger amount of deep processing and a smaller amount of shallow processing, and higher persistence than did framing tasks in terms of extrinsic future goals (e.g., the task is important for money or an attractive image in the future).

Finally, in a comprehensive review of the intrinsic-extrinsic future goal content literature relating specifically to academic outcomes, Vansteenkiste, Lens, and Deci (2006) commented that despite the fact that "only very recently have these differential [intrinsic-extrinsic future] goal contents been linked to academically relevant outcomes" (p. 23), evidence increasingly points in the direction indicating that intrinsic future goal content may be related to adaptive academic outcomes, whereas extrinsic future goal content may be related to maladaptive academic outcomes. Based on the research in this area, we decided to include a

distinction between intrinsic and extrinsic future goals in our test of predictions from the Miller and Brickman (2004) model.

Miller and Brickman (2004) viewed proximal subgoals as possibly including both "target subgoals," or middle-range intermediate subgoals of the kind described by Harackiewicz and Elliot (1998), as well as related, more-close-range sub-subgoals. The present study focused on one likely middle-range target subgoal for college students, namely, college graduation. The reasons for the choice of college graduation as the middle-range target subgoal of interest in the present study were its relatively high importance in a college environment and the fact that the college graduation subgoal was central to the measures of instrumentality and the self-regulation strategies that were tested.

Another important variable in the Miller and Brickman (2004) model is perceived task instrumentality. Perceived task instrumentality was defined as the perception that work on academic tasks

<sup>1</sup> The theoretical basis for Kasser & Ryan's (1993, 1996) classification of aspirations into "intrinsic" and "extrinsic" was the a priori premise in self-determination theory that the pursuit of goals that satisfies the theorized basic needs of competence, autonomy, and relatedness embodies "authentic," or self-determined, motivation, whereas the pursuit of goals emanating from external pressure (e.g., wealth, fame, and image) exemplifies externally controlled motivation (Ryan & Deci, 2000).

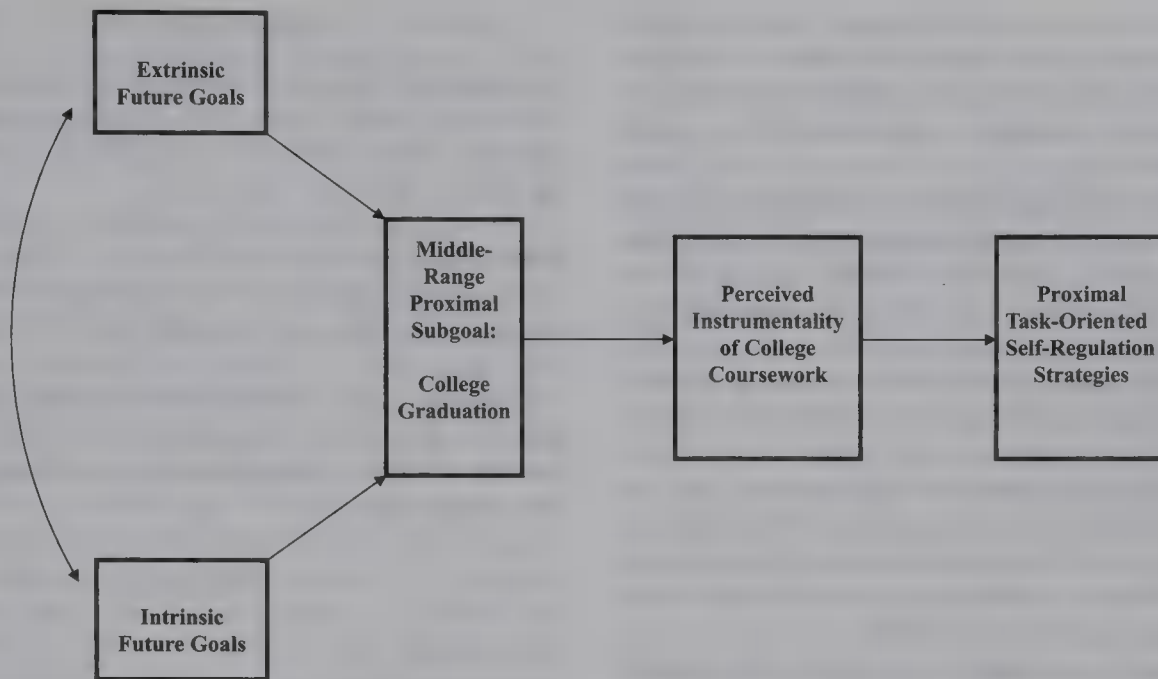


Figure 2. The theoretical model. Extrinsic and intrinsic future goals predict a middle-range subgoal, namely, the college graduation subgoal. The college graduation subgoal, in turn, predicts perceived task instrumentality. Finally, perceived task instrumentality predicts proximal task-oriented self-regulation strategies. The extrinsic and intrinsic future goals correlate.

(i.e., academic course work) is instrumental to one's future.<sup>2</sup> Miller and Brickman (2004) argued that it is the students' perceptions of task instrumentality that transmit the value of their future goals and subgoals to the proximal tasks they are faced with (Miller, DeBacker, & Greene, 1999). They further argued that students are more likely to perceive proximal tasks as instrumental when they have an adaptive system of proximal subgoals leading to personally valued future goals.

Finally, when proximal tasks are perceived to be instrumental, students are more likely to engage in proximal task-oriented self-regulation strategies to accomplish those tasks (e.g., Brickman & Miller, 2001; Greene, Miller, Crowson, Duke, & Akey, 2004; Miller, Greene, Montalvo, Ravindran, & Nichols, 1996). This perspective is in line with Bandura's (1986) observation that personal development is enhanced considerably when individuals connect "distal aspirations with proximal self-guidance" (p. 476). The variables of interest to the present study, along with the theorized relationships among them (Miller & Brickman, 2004), are summarized in the theoretical model (see Figure 2).

Although many of the relationships among the variables depicted in Miller and Brickman's (2004) work are based on research results, some relationships are based mainly on theoretical grounds, as research that specifically connects distal future goals and proximal subgoals is scarce (Husman & Lens, 1999; Locke & Latham, 1990), especially in educational environments. However, the few studies about the relationships between relatively distal goals and relatively proximal subgoals carried out in educational institutions make Miller and Brickman's (2004) hypothesis plausible. For example, in two studies conducted in a college and a high school environment, respectively, Schutz and Lanehart (1994) and Schutz (1997) found, among other things, that the students' distal educational goals (e.g., earn a Master's degree, earn a Doctorate) predicted their proximal educational subgoals

(e.g., "I read textbooks assigned for my class"), and that the distal educational goals indirectly predicted (through the educational subgoals and other variables) the students' grade point averages. Although the distal educational goals tapped by these two studies differed considerably from the personally valued future goals that were of interest in the present study, the findings were in line with Miller and Brickman's (2004) model. Finally, in a study based on an earlier formulation of their motivational model, Brickman and Miller (2001) conducted a qualitative study at an alternative high school and found significant relationships among sociocultural factors and future goals, subgoals, and perceptions of ability among the student participants. These factors were, in turn, related to perceptions of instrumentality of school tasks and to the proximal achievement goals reported. In addition, perceptions of instrumentality were related to self-reports of self-regulation and cognitive engagement and to specific observed patterns of task engagement. Taken together, the results of these studies lend support to Miller and Brickman's (2004) model.

### The Present Study

The purpose of the present study was to test the directional predictions of the portion of the Miller and Brickman (2004) model depicted in the theoretical model (see Figure 2). This portion of the model indicates that students' future goals predict

<sup>2</sup> The construct of perceived task instrumentality is clearly related to Eccles's (1983; Wigfield & Eccles, 2001) "utility value"; however, the constructs differ in their points of emphasis. Utility value refers to an individual's belief about the value of a task or goal object for accomplishing some other end. Miller and Brickman (2004) argued that it is the perception of the instrumentality of a task to future goal attainment that gives the task its value.



their adoption of a college graduation subgoal; that the college graduation subgoal predicts the students' perceptions of task instrumentality; and that the latter, in turn, predicts proximal task-oriented self-regulation strategies (also referred to as "self-regulation strategies").<sup>3</sup>

The present research was important to undertake for a number of reasons. First, it adds to the scarce research on the relationships among future and proximal motivation variables in academic environments, and thus deepens our comprehensive understanding of various facets of motivation. Second, it provides a test of a relatively new, integrated, and directional model of future and proximal motivation and self-regulation (Miller & Brickman, 2004). This study will either provide evidence for the model's hypotheses or suggest ways in which the model could be improved. Third, the current research examines Miller and Brickman's (2004) personally valued future goals differentially, in terms of their extrinsic and intrinsic hypothesized dimensions (Kasser & Ryan, 1993, 1996), an examination not carried out before.

Fourth, the present research examines factors thought to predict perceived task instrumentality. In educational environments, task instrumentality, or the perception among students that the tasks available in school are instrumental to their future, has been found to predict many positive and adaptive learning outcomes (De Volder & Lens, 1982; Greene et al., 2004; Malka & Covington, 2005; Miller et al., 1996; Raynor, 1970, 1974; Simons et al., 2004). Yet very few studies have attempted to answer the question of what predicts task instrumentality. Miller and Brickman (2004) hypothesized that when students have strong, self-determined distal future goals and related proximal subgoals, there is an increased likelihood that students will perceive relevant school tasks as instrumental to goal accomplishment. This study is among the very few to use future goals and a middle-range college graduation subgoal as predictors of task instrumentality. Finally, although (as noted above) task instrumentality has been shown to predict many adaptive educational outcomes, including self-regulation strategies, this study will examine the relationship between task instrumentality and task-oriented self-regulation strategies in a new context of a directional model consisting of interconnected future and proximal motivational variables.

The present study is based on a number of premises. For example, it is important to note that there is no accepted standard in the motivation literature by which a goal can be definitively considered a "future" or a "proximal" goal or subgoal. All goals are, to some extent, future-oriented, and researchers have investigated the relationships between people's *relatively* distal and *relatively* proximal goals and subgoals. In the present study, the most distal goals were the goals assessed by Kasser and Ryan's (2004) Aspirations Index, which tapped intrinsic and extrinsic life goals (e.g., contribution to community, personal wealth). These were referred to in the study as "future goals." A somewhat more middle-range proximal subgoal for these aspirations was that of graduating from college. Graduating from college could be considered for some a step to attaining one or more of the life future goals/aspirations previously mentioned. In the study, this subgoal was referred to as the "college graduation subgoal." The scales used to measure the distal goals and the college graduation subgoal will be described more fully in the Method section, along with all the other scales used in this study.

## Research Questions

Based on Miller and Brickman's (2004) hypothesis about the relationships among future goals, proximal subgoals, task instrumentality, and self-regulation strategies (see Figure 2), the following research questions were asked:

Does the theoretical model depicted in Figure 2 provide an adequate fit of the data?

Do intrinsic and extrinsic future goals have differential relationships with the other variables in the theoretical model (see Figure 2)?

If the theoretical model depicted in Figure 2 provides an adequate fit of the data, how confident can we be that the model provides a plausible variable and path configuration, rather than just a chance fit?

## Method

### Participants

Participants were 421 student volunteers enrolled in 18 sections of a 2nd-year English course, open to all university students and titled Literary Heritage, at a large, southern, urban university. Students in all the sections of this course whose instructors gave permission were asked to volunteer (out of 26 regular on-campus sections, teachers in 18 sections gave their permission for the study). In each section, the maximum enrollment was 35 students. Originally, a total of 422 students volunteered to participate; 1 student, however, filled out the Scantron answer sheet incorrectly and was dropped from the study. The demographic description of the students who were retained in the present study ( $N = 421$ ) was as follows: 88.6% were enrolled as full-time students, 11.4% were enrolled as part-time students, 52.5% were men, 47.5% were women, 13.5% were Black, 75.8% were White, .7% were American Indian/Alaska native, 3.1% were Asian, 1.2% were Mexican-American/Chicano, 1.4% were other Latino, and the rest were "other." Two participants did not report race. ACT scores were as follows: 3.3% reported ACT scores between 11 and 15 (or SAT scores between 500 and 750), 11.9% reported scores between 16 and 18 (SAT scores between 760 and 890), 26.8% reported scores between 19 and 21 (SAT scores between 900 and 1,010), 23% reported scores between 22 and 24 (SAT scores between 1,020 and 1,120), 21.4% reported scores between 25 and 27 (SAT scores between 1,130 and 1,230), and the rest (9.8%) reported scores greater than 27 (SAT scores greater than 1,230). Some students (3.8%) failed to report test scores, and personal discussions with teachers revealed that under special circumstances, students may be accepted without these standardized test scores.

### Procedure

The study was conducted in the students' classrooms. Potential participants were informed of the general nature of the study, as well as of possible adverse effects, and they were told that anonymity and confidentiality would be strictly maintained. Students were offered a small incentive in the form of a candy bar for their

<sup>3</sup> The term "predict" (and its derivatives) in the present study refers to shared variance between the pertinent variables.

participation. After signing a consent form, students were asked to complete the Future-Oriented Student Motivation Survey (FOSS), an instrument containing demographic information as well as four scales, as detailed below.

### Measures

*The Future-Oriented Student Motivation Survey (FOSS).* Participants were administered the FOSS (see sample items in Table 1). This instrument included a short demographic portion as well as four scales measuring different aspects of future- and proximally oriented motivation and self-regulation, as described below.

*The Future Goals Scale.* To measure the strength of personally valued future goals, the Aspirations Index (Kasser & Ryan, 2004) was used in a modified form. The Aspirations Index was chosen for the measurement of personal future goals because it includes seven major (and inclusive) aspirations, and it conceptually divides them into extrinsic and intrinsic aspirations. This allowed for more complexity in the data analysis and in our understanding of whether the intrinsic versus extrinsic nature of the aspirations are predictive of the other variables.

The 2004 version of the Aspirations Index (Kasser & Ryan, 2004) had a 7-point Likert-type scale. It included 35 aspirations representing seven life domains (each life domain being represented by five items). The extrinsic subscale included the life domains of wealth, attractive image, and fame, and the intrinsic subscale included the life domains of health, personal growth, affiliation, and community contribution. For each item representing a goal, three questions were asked: a value question (how important the goal is), an expectancy question (to what extent one expects to accomplish it), and an indirect commitment question (how much of the goal one has accomplished already).

In the present study, two modifications were made to the Aspirations Index: The expectancy question was dropped entirely, since it was not of interest in this study, and the indirect commitment question was changed to a direct commitment question (i.e., how committed one is to reaching the goal). Thus, in the present study, each aspiration item was followed by two questions: a value question and a direct commitment question. Based on the Miller and Brickman (2004) model and social-cognitive theory (Bandura, 1986, pp. 323, 477; Locke & Latham, 1990, p. 124), the value of, and commitment to, goals were thought to be among the most

important goal aspects. Also, the change from an indirect to a direct commitment question reflects the present study's underlying assumption, in line with social-cognitive theory (see Locke & Latham, 1990, p. 5), that people are aware, to a large extent, of their goal commitment levels. The participants responded on a 7-point scale, with 1 denoting *not at all* and 7 denoting *very*. The value and commitment questions were combined and averaged for each of the seven life domains.

*The College Graduation Subgoal Scale.* This scale was developed specifically for the present study and was designed to measure a plausible middle-range mediating subgoal between the distal future goals (as measured by the modified Aspirations Index) and students' perceptions of task instrumentality (i.e., their viewing of their college course work as instrumental for their goal attainment). The scale was modeled after the modified Aspirations Index described above. Following pilot testing, the initial five-item scale was reduced to three items. Each item asked the same two questions that were asked in the modified Aspirations Index: one about value, and one about commitment, for a total of six questions. The participants responded on a 7-point scale, 1 denoting *not at all* and 7 denoting *very*. The responses for all the questions were averaged. The alpha reliability indicated by a pilot study was .97.

*The Perceived Instrumentality Scale.* The Perceived Instrumentality Scale used by Greene et al. (2004) was selected because it was designed to measure instrumentality for school work, which was the variable of interest in this study. Both Greene et al. (2004) and Miller et al. (1999) provided convincing evidence for the reliability and validity of the scale. The scale contains five items measuring perceptions of instrumentality for school work (e.g., "I do the work assigned in this class because my achievement plays a role in reaching my future goal"). Participants responded on a 7-point scale, with 1 indicating *strongly disagree* and 7 indicating *strongly agree*.

*The Task-Oriented Self-Regulation Strategy Scale.* The extent of proximal task-oriented self regulation strategies was measured by a subset of the Motivated Strategies for Learning Questionnaire (Pintrich, Smith, Garcia, & McKeachie, 1991) that included eight learning strategy subscales conducive to college graduation. The eight Motivated Strategies for Learning Questionnaire learning strategy subscales we used were Rehearsal, Elaboration, Organization, Critical Thinking, Metacognitive Self-Regulation, Time

Table 1  
Scales Used in the Present Study With Sample Items and Measures of Central Tendency, Normality, and Reliability

Scale	Sample item(s)	M	SD	Skew	Kurtosis	$\alpha$ rel.
Future Goals Scale <sup>a</sup>	Extrinsic life goal: To be a very wealthy person.	3.95	1.09	-0.04	-0.52	.95
	Intrinsic life goal: To grow and learn new things.	5.85	0.64	-0.78	1.25	.93
College Graduation Subgoal Scale	Goal: To graduate from college.	6.68	0.66	-2.94	10.23	.88
Perceived Instrumentality Scale	I do my course work this semester because. . . My achievement plays a role in reaching my future goals.	5.72	1.33	-1.18	1.14	.92
Task-Oriented Self-Regulation Strategy Scale <sup>b</sup>	I usually study in a place where I can concentrate on my course work.	4.38	0.80	-0.12	0.03	.93

Note. The standard error of skew and kurtosis was .12 and .24, respectively, in most scales.  $\alpha$  rel. = alpha reliability.

<sup>a</sup> This scale is a modified version of the Aspirations Index (Kasser & Ryan, 2004). <sup>b</sup> This scale is a subset of the Motivated Strategies for Learning Questionnaire (Pintrich et al., 1991).



and Study Environment, Peer Learning, and Help Seeking. Participants responded to these items on a 7-point scale, with 1 indicating *not at all true of me* and 7 indicating *very true of me*.

## Results

### *Preliminary Data Analysis*

Prior to conducting the path analysis, we performed a preliminary analysis of the data to gauge whether the data were appropriate for use in a path analysis. A number of two-way contingency table analyses were conducted to evaluate whether the missing values were missing at random, and thus whether they were proportionally distributed by gender, age, student status (part time or full time), race, and expected grade point average. The evidence suggested that the missing data were missing at random. In addition, a before-and-after-mean-substitution examination of instrument reliabilities, measures of central tendency and normality, and correlation matrixes revealed minimal differences. Based on the nature of the data and the results of the preliminary analyses, we used mean replacement for missing values.

**Instrument reliabilities.** Cronbach alpha reliabilities were computed for all scales and subscales on the FOSS instrument to gauge the internal consistency of the scales. A summary of the alpha reliabilities is listed in Table 1. The Future Goals subscales had alpha reliability values of .95 and .93 for the Extrinsic and Intrinsic subscales, respectively, and the alpha reliabilities of the College Graduation Subgoal Scale, the Perceived Instrumentality Scale, and the Task-Oriented Self-Regulation Strategy Scale were .88, .92, and .93, respectively. The reliabilities for the scales in this study were deemed adequate for the present study.

**Measures of central tendencies and normality.** Descriptive information about the variables can be seen in Table 1. The variable with the most pronounced skew and kurtosis was the college graduation subgoal (skew of  $-2.94$ , kurtosis of  $10.23$ ), and that was expected in view of the fact that most college students were likely to indicate a strong desire to graduate from college. Other variables were slightly skewed or kurtotic, as can be seen in Table 1, but their deviation from normality was of small magnitude. In view of the expected deviation from normality in variables related to college graduation goals among the college student population sampled, the variables were left untransformed (see Ullman, 1996, p. 790, for a similar opinion).

For the present path analytic study, we used the maximum likelihood (ML) estimation method. Although the ML estimation is considered fairly robust against small-to-moderate violations of normality (Anderson & Gerbing, 1988; Jöreskog & Sörbom, 1989; McDonald & Ho, 2002), moderate-to-major violations can adversely affect the chi-square statistic that serves as an important measure of model fit. However, the particular way in which the chi-square is affected under conditions of data nonnormality was considered to render the ML estimation method suitable for the present study. According to Curran, West, and Finch (1996), there are two concerns about using the chi-square statistic under conditions of nonnormality: (a) a model might be mistakenly rejected when it is correct, and (b) a model might be opportunistically modified until an acceptable chi-square level is achieved, even though the model might be basically correct and in no need of modification.

In the present study, these two concerns were addressed in a number of ways. First, model fit decisions were not based solely on the chi-square statistic but on other goodness-of-fit indices (GFIs) as well, such as the GFI, the normed fit index (NFI), the nonnormed fit index (NNFI), the comparative fit index (CFI), and the root mean square error of approximation (RMSEA), all of which were reported in the present study.<sup>4</sup> Second, any consideration to remove or add a path was based on strong theoretical grounds rather than on chance trial-and-error opportunities. For example, such considerations were aided by the model parameter estimates, which have been found to be unbiased even under conditions of nonnormality (Curran et al., 1996; Enders, 2001; McDonald & Ho, 2002). Third, care was taken to limit model modifications to very few modifications to prevent "capitalization on chance characteristics of the data" (MacCallum, Roznowski, & Necowitz, 1992, p. 490). Finally, the path analysis was followed by a bootstrap resampling (replication) technique so we could investigate the stability and generalizability of our confirmatory factor analysis model (Efron & Tibshirani, 1998).

**Correlations matrix.** Bivariate scatterplots conducted on pairs of variables chosen at random indicated linear relationships in the data. No curvilinear relationships were observed. Table 2 presents the Pearson moment correlations among the variables of interest. The correlations were very consistent with the Miller and Brickman (2004) model. Nearly all the variables were significantly intercorrelated, and some of the relationships seemed stronger than others, mostly in expected directions. Following the theoretical predictions of Miller and Brickman (2004), future goals correlated significantly with the subgoal of college graduation. At the same time, in line with self-determination theory (Ryan & Deci, 2000), the relationship between extrinsic future goals and the college graduation subgoal ( $r = .12$ ;  $p < .05$ ) seemed to be weaker than the relationship between intrinsic future goals and the college graduation subgoal ( $r = .39$ ,  $p < .01$ ).

As theorized by Miller and Brickman (2004), the college graduation subgoal correlated with task instrumentality ( $r = .42$ ;  $p < .01$ ) as well as with the self-regulation strategies ( $r = .37$ ;  $p < .01$ ), and the self-regulation strategies were highly correlated with task instrumentality ( $r = .58$ ;  $p < .01$ ). In addition, extrinsic and intrinsic future goals correlated with task instrumentality ( $r = .17$  and  $r = .46$ , respectively;  $p < .01$ ), as well as with the self-regulation strategies ( $r = .21$  and  $r = .52$ , respectively;  $p < .01$ ). Again, extrinsic and intrinsic future goals seemed to have a differential relationship to task instrumentality and to self-regulation.

On the other hand, also consistent with self-determination theory and with Kasser and Ryan (1993, 1996), extrinsic and intrinsic goals were not mutually exclusive, and they correlated with each other. Extrinsic future goals correlated with intrinsic future goals ( $r = .33$ ;  $p < .01$ ).

<sup>4</sup> Most fit indexes have strengths and drawbacks. For example, Ullman (1996) reported that the NNFI may underestimate fit in samples with small numbers, although it is not clear how small. In this study, all four indexes were consulted, including the RMSEA indicator, before making judgments regarding model fit.

Table 2  
Correlations Among the Variables in the Present Study

Variable	1	2	3	4	5
1. Extrinsic future goals	1				
2. Intrinsic future goals	.33**	1			
3. College graduation subgoal	.12*	.39**	1		
4. Perceived task instrumentality	.17**	.46**	.42**	1	
5. Task-oriented self-regulation strategies	.21**	.52**	.37**	.58**	1

\*  $p < .05$ . \*\*  $p < .01$ .

### Model Tested

Path analysis was performed, using PROC CALIS (Hatcher, 1994) from SAS/STAT (information on this software can be found at <http://www.sas.com/technologies/analytics/statistics/stat/index.html>), to test the theoretical model (see Figure 2) describing the relationships among four major variables in Miller and Brickman's (2004) model. In the analyses, the ML method of parameter estimation was used, and all analyses were performed on the variance-covariance matrix ( $N = 421$  observations). In the theoretical model (see Figure 2), extrinsic future goals and intrinsic future goals (as measured by the Future Goals Scale) predict the subgoal of college graduation (as measured by the College Graduation Subgoal Scale), which, in turn, predicts task instrumentality (as measured by the Perceived Instrumentality Scale). Finally, perceived task instrumentality predicts task-oriented self-regulation strategies (as measured by the Task-Oriented Self-Regulation Strategy Scale). The extrinsic and intrinsic future goals are modeled as correlating, based on the significant correlations between them in the pilot study and in the present study. Theoretically, as well, these future goals, although shown to lead to differential well-being outcomes, have not been seen as mutually exclusive (e.g., Kasser & Ryan, 1993, 1996).

GFI for the theoretical model, the revised model, and the final model are presented in Table 3, and the bootstrap goodness-of-fit estimates based on the final model (see Figure 3) are presented in Table 4. The chi-square statistic included in Tables 3 and 4 provides a test of the null hypothesis that the reproduced covariance matrix has the specified model structure—in other words, that the model fits the data. Tables 3 and 4 also provide four additional GFIs: the GFI, the NFI (Bentler & Bonett, 1980), the NNFI (Bentler & Bonett, 1980), the CFI (Ullman, 1996), and the RMSEA (Byrne, 2001). GFI, NFI, NNFI, and CFI values of more than .9, and RMSEA values of less than .05 are generally thought to indicate a good fit between the model and the data.

The "null model" in Table 3 represents a hypothetical path model in which none of the variables are related to any of the other variables. This null model chi-square is useful as a baseline against which the chi-square values obtained for the other models can be compared. If the theoretical model achieves a large reduction in chi-square in comparison with the null model (while considering the degrees of freedom), then the theoretical model gains support.

**Estimation of the theoretical model.** Estimation of the theoretical model revealed a significant model chi-square value,  $\chi^2(5, N = 421) = 115.03, p < .001$ , indicating that the observed and

model-implied covariance matrices may be significantly different. Although the value of the GFI was an acceptable .916, the values of the NFI, CFI, and NNFI were .760, .531, and .765, respectively, much lower than the desired  $>.9$ , and the value of the RMSEA was .229, much higher than the desired  $<.05$  value. Taken together, these values indicated that the fit between the model and data could probably be improved.

The path coefficients in the theoretical model (see Table 5) were reviewed to see whether any paths should be deleted or added to improve model fit. The  $t$  values for most path coefficients proved to be statistically significant ( $p < .001$ ), with most  $t$  values exceeding 6.25.<sup>5</sup> Most standardized path coefficients were either equal to or exceeded .28 in absolute magnitude. One path, however, was not significant, namely, the path predicting the college graduation subgoal from the extrinsic future goals (standardized coefficient =  $-.004, t = -.08$ ).

Despite the fact that this path did not reach statistical significance, we decided to leave it in place. This decision was based on three major considerations: First, it was theoretically conceivable that the Future Goals Scale did not capture all the possible extrinsic future goals that people might have. Second, in the literature (e.g., Kasser & Ryan, 1993, 1996), intrinsic and extrinsic goals tend to correlate, albeit at relatively small magnitudes between  $r = .2$  and  $r = .3$ , indicating that people might operate with a mix of goals. Third, we wanted to see what the bootstrap technique would show about all these paths following 200 iterations of the model.

A careful examination of the parameter values in the model, along with a reexamination of the correlations table (see Table 2) and of other theoretical considerations, resulted in the decision to add a direct path predicting perceived task instrumentality from intrinsic future goals. The reasons for adding the path were at least threefold. First, in line with self-determination theory (Deci & Ryan, 2000), the theoretical model parameters indicated that intrinsic, rather than extrinsic, future goals were the ones directly predictive of the college graduation subgoal and, through this subgoal, of perceived task instrumentality. Second, the correlations table (Table 2) indicated a strong and significant relationship between intrinsic future goals and perceived task instrumentality ( $r = .46$ , in comparison with  $r = .17$  for the relationship between extrinsic future goals and task instrumentality). Finally, because it was conceivable (likely) that our subgoal measure did not capture all the variance attributable to intrinsic future goals, we believed that adding a direct path leading from intrinsic future goals to perceived task instrumentality might capture additional variance and improve model fit.

**Estimation of the revised model.** After adding a path predicting perceived task instrumentality from intrinsic future goals, the estimation of the revised model revealed that the hypothesized model, although improved over the theoretical model, still did not fit the data adequately (see Table 3). The chi-square was significant,  $\chi^2(4, N = 421) = 59.12, p < .001$ , indicating that the observed and model-implied covariance matrices were again significantly different. The values of all the indices improved, with the GFI, NFI, NNFI, and CFI at .950, .877, .706, and .882,

<sup>5</sup> These  $t$  tests are statistically significant at the  $p < .05$  level whenever their absolute value exceeds 1.96, at the  $<.01$  level if  $t$  exceeds 2.58, and at the  $<.001$  level if  $t$  exceeds 3.30 (two-tailed tests).



Table 3  
Goodness-of-Fit Indices for the Theoretical, Revised, and Final Models

Model	$\chi^2$	df	p	GFI	NFI	NNFI	CFI	RMSEA
Null model	478.70	10	<.001	0.000				
Theoretical model	115.03	5	<.001	.916	.760	.531	.765	.229
Revised model	59.12	4	<.001	.950	.877	.706	.882	.181
Final model	4.88	3	.181	.995	.990	.987	.996	.039

Note.  $N = 421$  participants. The revised model is identical to the theoretical model, except that a path was added from intrinsic future goals to perceived task instrumentality. The final model is identical to the revised model, except that a path was added from intrinsic future goals to task-oriented self-regulation strategies. GFI = goodness-of-fit index; NFI = normed fit index; NNFI = nonnormed fit index; CFI = comparative fit index; RMSEA = root mean square error of approximation.

respectively. The RMSEA also improved and was .181. Despite the improvement, however, these various measures indicated that the fit was still not adequate. We decided to revise the model again.

Based on similar theoretical rationales that guided the first model revision, we decided to add an additional path predicting task-oriented self-regulation strategies from the intrinsic future goals. The reasons for adding the path were, again, at least three-fold. First, in line with self-determination theory (Deci & Ryan, 2000), the theoretical model parameters indicated that intrinsic, rather than extrinsic, future goals were the ones directly predictive of the college graduation subgoal and, through this subgoal, of perceived task instrumentality and of task-oriented self-regulation strategies. Second, the correlations table (Table 2) indicated a strong significant correlation between intrinsic future goals and

task-oriented self-regulation strategies ( $r = .52$ , in comparison with  $r = .21$  for the relationship between extrinsic future goals and self-regulation strategies). Finally, as it was conceivable (likely) that our perceived instrumentality measure did not capture all the possible variance attributable to intrinsic future goals, we believed adding a direct path leading from intrinsic future goals to task-oriented self-regulation strategies might capture additional variance in self-regulation and improve model fit.

*Estimation of the revised model—2.* Following the second revision of the original theoretical model, the model fit the data quite well (see Table 3). The chi-square was not significant,  $\chi^2(3, N = 421) = 4.88, p = .181$ , indicating that there was no significant difference between the observed and the model-implied covariance matrices. The values of all the fit indices were over .98, as follows: GFI = .995, NFI = .990, NNFI = .987, and CFI = .996. The

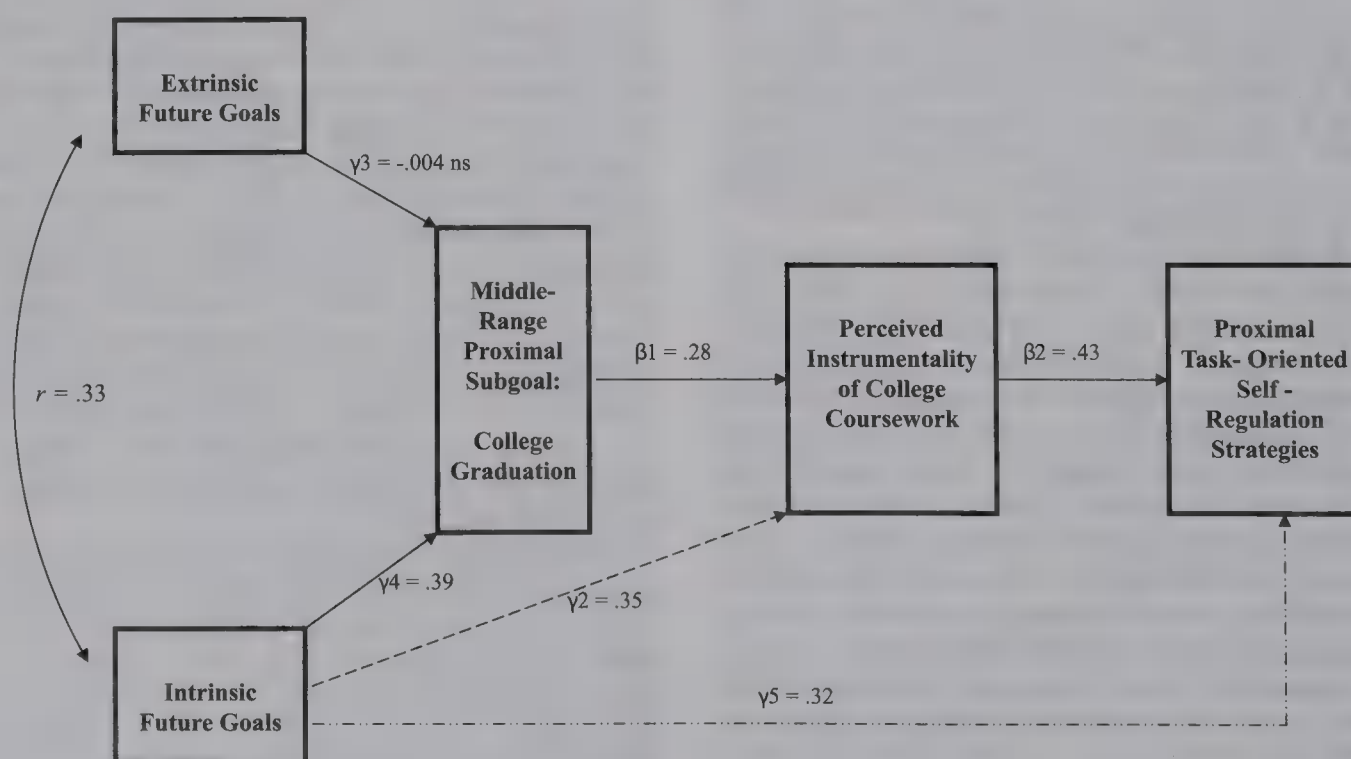


Figure 3. The final model. The broken lines indicate two paths that were added during two subsequent revisions of the theoretical model. These additional paths, predicting perceived task instrumentality from intrinsic future goals and predicting task-oriented self-regulation strategies from intrinsic future goals, constitute the only difference between the theoretical model and the final model. The numbers indicate standardized path analysis coefficients (betas and gammas), with the exception of  $r$ , which indicates the Pearson product-moment correlation between intrinsic and extrinsic future goals. The letters ns indicate that a path is nonsignificant. All significant standardized coefficients had significant  $ts$  at the  $p < .001$  level.

Table 4  
*Bootstrap Goodness-of-Fit Indices for Final Model*

Statistic or index	<i>M</i>	<i>SD</i>	25%	50%	75%	Min., max.
$\chi^2$	7.889	5.681	3.947	6.546	10.715	0.103, 35.237
<i>df</i>	3	3	3	3	3	3
<i>p</i>	.189 <sup>a</sup>	.239	.267 <sup>b</sup>	.088 <sup>b</sup>	.013 <sup>b</sup>	<.001, .991
GFI	.993	.005	.990	.994	.996	.969, .999
NFI	.984	.011	.978	.987	.992	.930, .999
NNFI	.966	.038	.947	.976	.993	.783, 1.022
CFI	.989	.011	.984	.993	.998	.935, 1.000
RMSEA	.053	.036	.027	.053	.078	<.001, .160

*Note.* Bootstrap analysis was based on 200 iterations. Min. = minimum value; max. = maximum value. GFI = goodness-of-fit index; NFI = normed fit index; NNFI = nonnormed fit index; CFI = comparative fit index; RMSEA = root mean square error of approximation.

<sup>a</sup> Average of all probabilities across 200 iterations. <sup>b</sup> Percentile values established by ordering probabilities from high to low values. Values are equivalent to calculating probabilities from the average percentile chi-squares with *df* = 3.

RMSEA was .039, clearly within the <.05 optimal level. It was decided to accept this model as the final model.

In the final model (see Table 5), the path values indicated that task instrumentality was significantly and positively predicted by the college graduation subgoal (standardized coefficient = .28, *t* = 6.25) and by intrinsic future goals (standardized coefficient = .35, *t* = 7.73). Self-regulation strategies were predicted by task instrumentality (standardized coefficient = .43, *t* = 10.19) and by intrinsic future goals (standardized coefficient = .32, *t* = 7.61). Finally, the college graduation subgoal was predicted by intrinsic future goals (standardized coefficient = .39, *t* = 8.10) but not by extrinsic future goals (standardized coefficient = -.004, *t* = -.08, *ns*). Together, the predictors in the final model (see Figure 3) accounted for approximately 84% of the variance in the predicted variables in the model.

### *Bootstrap Analysis*

To investigate the stability and generalizability of our model, we used a bootstrap resampling or, to be more precise, replication technique (Efron & Tibshirani, 1998). In this process we created *B* = 200 new data sets from the original data set and inspected the distribution of several fit measurements.<sup>6</sup> Each new data set contained *N* = 421 observations, and each observation was sampled with a replacement from the original data set. The results of the bootstrap analysis for the final model are presented in detail in Tables 4 and 5.

**Bootstrap model estimation.** The bootstrap analysis based on the final model (see Figure 3) showed evidence of good fit and remarkable stability across the 200 iterations. Table 4 summarizes the mean estimate of the 200 iterations for various model fit indicators as follows: At 3 *df*, the mean chi-square was 7.889, and the mean probability was .189, indicating good fit. Other indicators also showed evidence of good fit, such as the mean GFI (.993), NFI (.984), NNFI (.966), and CFI (.989). The mean RMSEA, slightly above the optimal <.05 at .053, still showed evidence of fair fit (Browne & Cudeck, 1993; MacCallum, Browne, & Sugawara, 1996).<sup>7</sup> Table 4 further shows the values of the various fit

indicators at the 25%, 50%, and 75% quartiles of the bootstrap iterations. At all of these points, most fit indicators showed evidence of good fit. For example, the GFI, NFI, NNFI, and CFI all had values above .94 at all the quartile points. The RMSEA values were between .027 and .053 at most of these quartile points, except for the 75% quartile, where the RMSEA value was .078, higher than the optimal <.05 but still within the fair fit guidelines (MacCallum et al., 1996; see also footnote 7). In addition, the standard deviation across the 200 iterations for most of the indicators was relatively small, indicating good model stability. For example, the standard deviation of the GFI, NFI, NNFI, CFI, and RMSEA fell between .005 and .038. The standard deviation of the chi-square was larger (5.681), and that of the probability was .239, but that was to be expected in view of the fact that some of the variables (notably the college graduation subgoal) deviated from normality. The chi-square statistic is known to be particularly sensitive to "departures from multivariate normality" (Ho, 2006, p. 285).

**Bootstrap parameter estimation.** The mean standard coefficients (betas and gammas) of the 200 bootstrap iterations tended to have nearly identical values to the paths in the final model (see Table 5), thus strengthening our confidence that the final model presents a very plausible explanation of the relationships among the variables of interest.

## Discussion

A directional model based on Miller and Brickman's (2004) model of future-oriented motivation and self-regulation was tested to find out whether the directional ordering of variables was supported by data in an academic setting and whether the predicted influence of future goals on proximal subgoals, task instrumentality, and self-regulation strategies would differ if the extrinsic and intrinsic nature of the future goals was considered. Although we think that our findings clearly address these issues, we want to strike some notes of caution before discussing the interpretations of our findings.

The correlational and predictive methods used in this study do not provide cause-and-effect evidence among the variables examined. Although these nonexperimental methods can provide evidence regarding the plausibility of the proposed paths, model fit in itself is not an indication that the data would not fit other types of relationships or variable configurations. There is a need for future research that continues to investigate the relationships among future goals, proximal subgoals, task instrumentality, and self-regulation strategies using different methodologies, and there is a need for additional experimental research that tests whether observed predictions follow a cause-and-effect pattern.

In addition, although the size of the sample in the present study was adequate for the study's purpose (*N* = 421), the sampling

<sup>6</sup> Efron and Tibshirani (1998) found 200 iterations as adequate in most cases.

<sup>7</sup> MacCallum et al. (1996, p. 134) summarized the RMSEA model fit guidelines, while taking into account Browne and Cudeck's (1993) guidelines, as follows: RMSEA values of less than .05 indicate close fit, values between .05 and .08 indicate fair fit, values between .08 and .10 indicate mediocre fit, and values above .10 indicate poor fit.



Table 5  
*Standardized Coefficients for the Paths in the Theoretical, Revised, and Final Models and in the Bootstrap Analysis of the Final Model*

Path	Standardized coefficient			Bootstrap analysis		
	Theoretical model	Revised model	Final model	Mean st. coeff.	SD	95% CI st. coeff.
Pred. self-regulation strategies from task instrumentality ( $\beta_2$ )	.57 <sup>a</sup>	.57 <sup>a</sup>	.43 <sup>a</sup>	.43 <sup>a</sup>	.03	.37, .48
Pred. self-regulation strategies from intrinsic future goals ( $\gamma_5$ )			.32 <sup>a</sup>	.32 <sup>a</sup>	.03	.27, .38
Pred. task instrumentality from the college graduation subgoal ( $\beta_1$ )	.42 <sup>a</sup>	.28 <sup>a</sup>	.28 <sup>a</sup>	.28 <sup>a</sup>	.05	.19, .36
Pred. task instrumentality from the intrinsic future goals ( $\gamma_2$ )		.35 <sup>a</sup>	.35 <sup>a</sup>	.35 <sup>a</sup>	.05	.26, .44
Pred. college graduation subgoal from extrinsic future goals ( $\gamma_3$ )	-.004, <i>ns</i>	-.004, <i>ns</i>	-.004, <i>ns</i>	.003, <i>ns</i>	.05	-.08, .09
Pred. college graduation subgoal from intrinsic future goals ( $\gamma_4$ )	.39 <sup>a</sup>	.39 <sup>a</sup>	.39 <sup>a</sup>	.38 <sup>a</sup>	.05	.30, .47
Correlation between extrinsic & intrinsic future goals	$r = .33^a$	$r = .33^a$	$r = .33^a$	$r = .33^a$	.04	.26, .40

*Note.* Bootstrap analysis was based on 200 iterations. *ns* denotes a non-significant *t*. The standardized coefficient columns indicate the standardized betas or gammas of the various paths, except for *r*, which indicates the Pearson product-moment correlation. St. coeff. = standardized coefficient; CI = confidence interval; Pred. = predicting.

<sup>a</sup>Significant *t* at the  $p < .001$  level.

method and the study design, involving a convenience sample and data collection during one limited period of time in one university department, limit the generalizability of the study. Additional studies will be needed before findings can be generalized to other disciplines, educational environments, and populations. For example, future research is needed in other disciplines such as business, law, engineering, and medicine to determine whether intrinsic future goals are still the most predictive of adaptive outcomes or whether extrinsic future goals play an important role as well. Additional research is needed at other levels of schooling, such as the elementary and high-schools levels, as well as in private and public schools. With these cautions in mind, we turn now to the discussion and interpretation of our findings.

### The Final Model

The theoretical model in the present study (see Figure 2) was modified by the addition of two paths, and the final model (see Figure 3) showed evidence of good fit with the data. A bootstrap procedure consisting of 200 iterations performed on the basis of the final model added evidence of the model's stability and generalizability.

Consistent with social-cognitive theory (Bandura, 1986) and with Miller and Brickman (2004), future goals had a significant direct and/or mediated relationship to the college graduation subgoal, the perceived task instrumentality, and the self-regulation strategies. Consistent with self-determination theory (Kasser & Ryan, 1993, 1996), the extrinsic and intrinsic nature of the students' future goals had a differential relationship to their subgoal of college graduation, their perceptions of task instrumentality, and their self-regulation strategies. Perceived task instrumentality was directly predicted by the college graduation subgoal and both directly and indirectly by the intrinsic future goals. Self-regulation strategies were directly predicted by task instrumentality and both directly and indirectly by intrinsic future goals. Finally, the college graduation subgoal was directly predicted by intrinsic, but not by extrinsic, future goals.

### Hypothesized Relationships Among the Variables of Interest

The present study's results indicate that Miller and Brickman's (2004) hypothesized mediated relationships among distal future goals, proximal subgoals, perceived task instrumentality, and task-oriented self-regulation are plausible ones: Future goals were found to predict the proximal subgoal of interest to this study, namely, college graduation, and, in turn, the college graduation subgoal was found to predict perceived task instrumentality. Finally, task instrumentality was found to predict task-oriented self-regulation strategies. The additional variance captured by two direct paths added to improve model fit (i.e., the path between future goals and task instrumentality and the path between future goals and task-oriented self-regulation strategies) was not surprising because the measures of the college graduation subgoal and perceived instrumentality used in the present study were unlikely to have accounted for all of the variation in these constructs attributable to intrinsic future goals.

The present study makes a contribution to the literature on goals by providing evidence that distal future goals are significantly related to proximal subgoals, to perceived task instrumentality, and to proximal task-oriented self-regulation, thus strengthening the argument that research on future-oriented and proximal motivation should be integrated (Husman & Lens, 1999; Kauffman & Husman, 2004). Also, the present study is among the first to test a specific hypothesis involving personally valued future goals and their predicted relationship to proximal subgoals, to task instrumentality, and to task-oriented self-regulation strategies in an educational setting. Results show a plausible directional path between the adoption of personally valued future goals, proximal subgoals, task instrumentality, and task-oriented self-regulation strategies at the college level.

### Task Instrumentality

The present study's contribution to our understanding of factors predicting task instrumentality deserves particular attention. Perceptions of task instrumentality have been found to have many

adaptive educational outcomes (e.g., Greene et al., 2004; Raynor, 1970, 1974; Vansteenkiste, Simons, Lens, Soenens, et al., 2004), including the adaptive task-oriented self-regulation strategies tested in the present study. Until recently, most studies of perceived task instrumentality have used task instrumentality as the predictor variable, which has made it hard to find out what factors may predict task instrumentality itself (e.g., Malka & Covington, 2005; Simons, Dewitte, & Lens, 2003). Because of the importance of perceptions of task instrumentality in educational environments, there has been a need to find factors that may predict task instrumentality.

Recently, Greene et al. (2004) found classroom-level context variables (perceived task meaningfulness, autonomy support, and mastery evaluation) to be important direct and indirect predictors of perceived task instrumentality. Earlier, Maehr and Midgley (1991) suggested that students' motivation regarding the tasks that they are asked to engage in may be influenced not only by classrooms and teachers, but also by school-level factors. In three case studies, Brickman and Miller (2001) found that students' past experiences in the larger sociocultural context (e.g., at home, at school, with peers, and through the media) were related to their perceptions of task instrumentality, possibly through the students' future goals, proximal subgoals, and perceptions of ability.

The present study found that the college graduation subgoal was a direct predictor of task instrumentality, and that intrinsic future goals were both direct and indirect (through the college graduation subgoal) predictors of task instrumentality. According to the directional model tested, the present study suggests that the students' own college graduation subgoal and future goals were the variables predicting perceived task instrumentality. Thus, to the possibility that task instrumentality may be predicted by community-, school-, and classroom-level factors, we may now add the possibility that task instrumentality may also be predicted by the students' own intrinsic future goals and by their own proximal subgoals, such as the college graduation subgoal.

### *Intrinsic Versus Extrinsic Future Goals*

The present study found that a focus on intrinsic, rather than extrinsic, future goals may be more predictive of the adoption of robust proximal subgoals, of adaptive perceptions of task instrumentality, and of the adoption of task-oriented self-regulation strategies in college environments. Kasser and Ryan (1993, 1996) have noted earlier that not all personally valued future goals were equally likely to predict similarly positive outcomes. These researchers found that extrinsic future goals were associated with diminished well-being in the present, while intrinsic future goals were associated with relatively high levels of well-being in the present. The present study is among the first to find evidence for a similar differential effect of extrinsic and intrinsic future goals in a different context than the studies mentioned above, namely, when future goals were used to predict proximal subgoals, task instrumentality, and task-oriented self-regulation strategies among college students.

Although this finding should be treated with caution until further evidence emerges, there is some indirect evidence that lends it additional support. Writing from a goal orientation theoretical perspective, Nicholls, Patashnick, and Nolen (1985) examined whether high school students' (relatively proximal) personal goals

in school were related to their perceptions of what the (relatively distal) aims of education should be. Despite the fact that Nicholls et al.'s (1985) construct of "aims of education" differed considerably from the present study's construct of personally valued future goals, their findings were in the same direction as those in the present study. The authors found that high school students who perceived the aim of education to be furthering one's wealth and status tended to have maladaptive personal school goals such as work avoidance, and they also tended to have ego, rather than task, orientations. On the other hand, students who perceived the aim of education to be commitment to society or understanding of the world tended to have adaptive personal school goals such as working hard, and they also tended to have task, rather than ego, orientations. These results are consistent with self-determination theory (Ryan & Deci, 2000).

In the present study, consistent with Kasser and Ryan (1993, 1996), personally valued intrinsic future goals were the pivotal point in predicting many positive factors, both directly and indirectly (e.g., college graduation subgoal, task instrumentality, task-oriented self-regulation strategies), whereas personally valued extrinsic future goals failed to produce a statistically significant relationship with the college graduation subgoal, the primary mediator for the rest of the model. The personally valued intrinsic future goals that the present study tested were individual growth, relationships, community involvement, and health. The personally valued extrinsic future goals tested were wealth, fame, and appearance.

### *Educational Implications*

Although the correlational nature of our study prohibits drawing causal conclusions, the clarity of the findings and the strong theoretical basis of the directional ordering underlying the model tested lead us to speculate about possible educational implications, should the model be supported by additional research using various methodologies (e.g., longitudinal studies) and by experimental research in particular. It is interesting to note that it may be the types of interventions hinted at below that provide experimental support for the model's validity. We see three areas with potential implications: the importance of students clarifying their future goals and subgoals, the utility of perceived instrumentality as a diagnostic tool for important motivation problems, and the importance of emphasizing the intrinsic goals of schooling. Each of these ideas is elaborated below.

*Clarification of personal goals and subgoals.* The high dropout rates in high schools and colleges in the United States, especially among poor students (National Center for Educational Statistics [NCES], 2004), raise the possibility that at least some of the students do not have an awareness of their own goals, have not done much thinking about aligning their future and proximal goals and subgoals in any coherent way, and have no idea where they are headed. The present research has indicated that the direction in which goals affect students may be from distal intrinsic future goals to proximal subgoals. Based on this knowledge, it may be possible to design goal-based interventions targeting at-risk students that would explain how a goal system works, help students identify their own long-term intrinsic goals, and set subgoals along the way, leading to the more distal future goals. For students with



no adaptive long-term intrinsic goals, it may be possible to design an intervention to foster such beneficial goals.

Students at risk of dropping out of school are often offered courses in remedial or study skills, and yet most college students enrolled in these remedial courses end up dropping out of school (NCES, 2004). The present study points to a possible reason. Study skills are the types of things that make up the self-regulation strategies that students normally set for themselves, such as study in a quiet place, study with a friend for a test, and summarize main ideas to oneself, among others. In the present study, these types of strategies were shown to be directly predicted by perceived instrumentality, directly and indirectly predicted by intrinsic future goals, and indirectly predicted by the college graduation subgoal. It may well be the case that students exhibiting problems in the self-regulation strategy area may have motivational problems that start with a lack of awareness of the larger goals at hand. Rather than focus on teaching a battery of standardized study skills, remedial programs may be more beneficial if they first helped students clarify or develop intrinsic future goals and proximal subgoals leading to their personal future goals, and if they then helped students perceive their work as instrumental toward achieving their goals. With their future goals and subgoals in place, and with well-developed perceptions of instrumentality, students may be in a much better position to act on improving their task-oriented self-regulation strategies or study skills.

*Perceived task instrumentality as an indicator.* The contribution of perceived task instrumentality to achievement and to other motivational factors in academic settings has been widely recognized (e.g., Brickman & Miller, 2001; Greene et al., 2004; Miller & Brickman, 2004; Raynor, 1970, 1974; Vansteenkiste, Simons, Lens, Soenens, et al., 2004). For example, Miller and Brickman (2004) as well as Greene et al. (2004) suggested that perceived task instrumentality may function as a helpful incentive when a student has to do school work that is not inherently pleasurable. The present study suggests that evident problems with perceived task instrumentality may serve as an early warning signal that can alert teachers and parents that their student may be having a more serious motivational problem. The present study found that the paths leading to perceptions of task instrumentality from intrinsic future goals and from the college graduation subgoal were significant. In other words, problems with task instrumentality (e.g., not seeing the reason why one should do a school assignment or thinking that all assignments are worthless) may be related to much larger problems, such as a lack of appropriate subgoals (e.g., school graduation), a lack of appropriate intrinsic future goals, or a combination of these factors. Accordingly, when a student exhibits signs of weak or nonexistent task instrumentality, educators should take it seriously and look beyond the specific assignment that was not turned in to identify other possible underlying problems. Such problems may include not only the student-level goal factors identified by the present study, but also classroom- and school-level factors identified by other studies (e.g., Brickman & Miller, 2001; Greene et al., 2004; Maehr & Midgley, 1991). In dealing with perceived task instrumentality problems on the part of students, educators may start by addressing student-level factors, such as discussing the student's intrinsic future goals and their subgoal or subgoals. If needed, educators may then widen their intervention to include classroom-, school-, and community-level factors, to the extent possible.

*Focus on intrinsic future goals.* The present study has found that the major predictor of a positive educational goal system and task instrumentality is the adoption of intrinsic, rather than extrinsic, future goals. Although these results need to be treated with caution until more evidence becomes available, it is interesting to note that, whereas some studies have identified possible drawbacks to extrinsically focused future goals (e.g., Kasser & Ryan, 1993, 1996; Nicholls et al., 1985; Ryan et al., 1999; Vansteenkiste, Simons, Lens, Soenens, et al., 2004), there are almost no studies identifying drawbacks to intrinsically focused future goals. In the present study, the paths from the intrinsic future goals to all other variables were significant, whereas the path from extrinsic future goals linking this variable to the rest of the mediated model was not significant. At the same time, intrinsic and extrinsic future goals were not mutually exclusive. In the two models tested, extrinsic and intrinsic future goals had a moderate correlation, implying a relationship. The implication may be that, although people may have a mix of extrinsic and intrinsic future goals, in order to be successful in an academic environment, they may need a stronger focus on intrinsic goals, such as personal growth, relationships, community involvement, and health.

For educators, these findings imply that they should encourage students to excel not by pointing out how much higher their salaries would be if they graduated from their respective schools, but by pointing out, for example, how much the students would know or how they might be able to contribute to society. The great importance accorded to educational improvement in the United States (e.g., United States Department of Education, 1983, 1994, 2001), coupled with the recognized potential of intrinsic future goals in particular, and of intrinsic factors in general, to improve schools, has already led many educators and researchers to design programs of school improvement that emphasize intrinsic elements (e.g., Brown, 1997; Huffman & Hipp, 2003; Mertens, Flowers, & Mulhall, 2001; Newmann & Wehlage, 1993; O'Hair & Odell, 1995; O'Hair & Reitzug, 1997; Raywid & Oshiyama, 2000; Wenger & Snyder, 2000).

## Conclusion

The purpose of the present study was to test a portion of the Miller and Brickman (2004) model of future-oriented motivation and self-regulation to find out whether students' distal future goals are related to their adoption of proximal subgoals, to their perceptions of task instrumentality, and to their task-oriented self-regulation strategies. The results supported Miller and Brickman's (2004) hypothesis that future goals predict the adoption of proximal subgoals, that proximal subgoals predict perceptions of task instrumentality, and that perceived instrumentality, in turn, predicts task-oriented self-regulation strategies. In line with self-determination theory (Kasser & Ryan, 1993, 1996), the present study also found that future goals had a differential relationship to the variables of interest: Although extrinsic future goals were not significant predictors of the college graduation subgoal, the primary mediator for the rest of the model, intrinsic future goals were significant direct predictors of the college graduation subgoal and direct and indirect predictors of perceptions of task instrumentality and of task-oriented self-regulation strategies. This finding suggests that, pending additional research, educators should consider emphasizing intrinsic future goals (and other intrinsic factors),



such as personal growth, meaningful relationships, and community contributions, in school environments to better facilitate the development of students' future goals, proximal subgoals, perceptions of task instrumentality, and task-oriented self-regulation strategies. In addition, the results of the present study indicate that the Miller and Brickman (2004) model of future-oriented motivation and self-regulation may serve as a basis for designing goal-based interventions to help at-risk students stay in school and succeed academically.

In their article about future-oriented motivation and self-regulation, Miller and Brickman (2004) urged that

those interested in proximal research issues and those with more future-oriented research agendas need to join forces in studying the phenomenon of academic motivation and self-regulation, and in planning interventions designed to improve the lives of the countless students who fail to see the relevance of schooling in their lives. (p. 29)

By finding a meaningful and significant connection between future goals, proximal subgoals, perceived task instrumentality, and task-oriented self-regulation strategies, the present study lends additional support to the need for continued attempts at integrating future-oriented and proximally oriented motivation and self-regulation.

## References

- Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103, 411–423.
- Bandura, A. (1986). *Social foundations of thought and action*. Englewood Cliffs, NJ: Prentice Hall.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.
- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588–606.
- Brickman, S. J., & Miller, R. B. (2001). The impact of sociocultural context on future goals and self-regulation. In D. M. McInerney & S. Van Etten (Eds.), *Research on sociocultural influences on motivation and learning* (pp. 119–137). Greenwich, CT: Information Age.
- Brown, A. L. (1997). Transforming schools into communities of thinking and learning about serious matters. *American Psychologist*, 52, 399–413.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.
- Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.
- Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, 1, 16–29.
- Deci, E. L., & Ryan, R. M. (2000). The “what” and “why” of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11, 227–268.
- De Volder, M. L., & Lens, W. (1982). Academic achievement and future time perspective as a cognitive-motivational concept. *Journal of Personality and Social Psychology*, 42, 566–571.
- Eccles, J. S. (1983). Expectancies, values and academic behaviors. In J. T. Spence (Ed.), *Achievement and achievement motives* (pp. 75–146). San Francisco, CA: Freeman.
- Efron, B., & Tibshirani, R. J. (1998). *An introduction to the bootstrap*. Boca Raton, FL: CRC Press.
- Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, 6, 352–370.
- Greene, B. A., Miller, R. B., Crowson, H. M., Duke, B. L., & Akey, K. L. (2004). Predicting high school students' cognitive engagement and achievement: Contributions of classroom perceptions and motivation. *Contemporary Educational Psychology*, 29, 462–482.
- Harackiewicz, J. M., & Elliot, A. J. (1998). The joint effects of target and purpose goals on intrinsic motivation: A mediational analysis. *Personality and Social Psychology Bulletin*, 24, 675–690.
- Hatcher, L. (1994). *A step-by-step approach to using SAS for factor analysis and structural equation modeling*. Cary, NC: SAS Institute, Inc.
- Ho, R. (2006). *Handbook of univariate and multivariate data analysis and interpretation with SPSS*. Boca Raton, FL: CRC Press.
- Huffman, J. B., & Hipp, K. K. (2003). *Reculturing schools as professional learning communities*. Lanham, MD: Scarecrow Education.
- Husman, J., & Lens, W. (1999). The role of the future in student motivation. *Educational Psychologist*, 34, 113–125.
- Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: A guide to the program and applications*. Chicago, IL: SPSS Inc.
- Kasser, T., & Ryan, R. M. (1993). A dark side of the American dream: Correlates of financial success as a central life aspiration. *Journal of Personality and Social Psychology*, 65, 410–422.
- Kasser, T., & Ryan, R. M. (1996). Further examining the American dream: Differential correlates of intrinsic and extrinsic goals. *Personality and Social Psychology Bulletin*, 22, 280–287.
- Kasser, T., & Ryan, R. M. (2004). *Aspirations Index*. Retrieved September 30, 2004, from the University of Rochester Self-Determination Theory website at [http://www.psych.rochester.edu/SDT/measures/aspir\\_scl.html](http://www.psych.rochester.edu/SDT/measures/aspir_scl.html)
- Kauffman, D. F., & Husman, J. (2004). Effects of time perspective on student motivation: Introduction to a special issue. *Educational Psychology Review*, 16, 1–7.
- Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting and task performance*. Englewood Cliffs, NJ: Prentice Hall.
- MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1, 130–149.
- MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, 111, 490–504.
- Maehr, M. L. (1984). Meaning and motivation: Toward a theory of personal investment. In R. Ames & C. Ames (Eds.), *Research on motivation in education: Student motivation* (Vol. 1, pp. 115–144). San Diego, CA: Academic Press.
- Maehr, M. L., & Midgley, C. (1991). Enhancing student motivation: A schoolwide approach. *Educational Psychologist*, 26, 399–427.
- Malka, A., & Covington, M. V. (2005). Perceiving school performance as instrumental to future goal attainment: Effects on graded performance. *Contemporary Educational Psychology*, 30, 60–80.
- McDonald, R. P., & Ho, M. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, 7, 64–82.
- Mertens, S. B., Flowers, N., & Mulhall, P. F. (2001). School size matters in interesting ways: Research on middle school renewal. *Middle School Journal*, 32, 51–55.
- Miller, R. B., & Brickman, S. J. (2004). A model of future-oriented motivation and self-regulation. *Educational Psychology Review*, 16, 9–33.
- Miller, R. B., DeBacker, T. K., & Greene, B. A. (1999). Perceived instrumentality and academics: The link to task valuing. *Journal of Instructional Psychology*, 26, 250–261.
- Miller, R. B., Greene, B. A., Montalvo, G. P., Ravindran, B., & Nichols,



- J. D. (1996). Engagement in academic work: The role of learning goals, future consequences, pleasing others, and perceived ability. *Contemporary Educational Psychology*, 21, 388–422.
- National Center for Educational Statistics. (2004). *The condition of education: Student effort and educational progress*. Retrieved September 30, 2004, from <http://nces.ed.gov/programs/coe/2004/section3/indicator16.asp> and <http://nces.ed.gov/programs/coe/2004/section3/indicator18.asp>
- Newmann, F. M., & Wehlage, G. G. (1993). Five standards of authentic instruction. *Educational Leadership*, 50, 8–12.
- Nicholls, J. G., Patashnick, M., & Nolen, S. B. (1985). Adolescents' theories of education. *Journal of Educational Psychology*, 77, 683–692.
- Nuttin, J. R. (with Lens, W.). (1985). *Future time perspective and motivation*. Leuven: Leuven University Press.
- O'Hair, M. J., & Odell, S. J. (Eds.). (1995). *Educating teachers for leadership and change*. Thousand Oaks, CA: Corwin Press.
- O'Hair, M. J., & Reitzug, U. C. (1997). Teacher leadership: In what ways? For what purposes? *Action in Teacher Education*, 19, 65–76.
- Pintrich, P. R., Smith, D. A., Garcia, T., & McKeachie, W. J. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)* (Report No. NCRIPAL-91-B-004). Ann Arbor, MI: National Center for Research to Improve Postsecondary Teaching and Learning. (ERIC Document Reproduction Service No. ED338122)
- Raynor, J. O. (1970). Relationships between achievement-related motives, future orientation, and academic performance. *Journal of Personality and Social Psychology*, 15, 28–33.
- Raynor, J. O. (1974). Future orientation in the study of achievement motivation. In J. W. Atkinson & J. O. Raynor (Eds.), *Motivation and achievement* (pp. 121–154). Washington, DC: Winston.
- Raynor, J. O., & Entin, E. E. (1982). Theory and research on future orientation and achievement motivation. In J. O. Raynor & E. E. Entin (Eds.), *Motivation, career striving, and aging* (pp. 13–82). New York: Hemisphere.
- Raywid, M. A., & Oshiyama, L. (2000). Musings in the wake of Columbine: What can schools do? *Phi Delta Kappan*, 81, 444–449.
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55, 68–78.
- Ryan, R. M., Little, T. D., Sheldon, K. M., Timoshina, E., & Deci, E. L. (1999). The American dream in Russia: Extrinsic aspirations and well-being in two cultures. *Personality and Social Psychology Bulletin*, 25, 1509–1524.
- Schutz, P. A. (1997). Educational goals, strategies use and the academic performance of high school students. *High School Journal*, 80, 193–202.
- Schutz, P. A., & Lanehart, S. J. (1994). Long-term educational goals, subgoals, learning strategies use and the academic performance of college students. *Learning & Individual Differences*, 6, 399–412.
- Simons, J., Dewitte, S., & Lens, W. (2000). Wanting to have versus wanting to be: The effect of perceived instrumentality on goal orientation. *British Journal of Psychology*, 91, 335–351.
- Simons, J., Dewitte, S., & Lens, W. (2003). "Don't do it for me. Do it for yourself!" Stressing the personal relevance enhances motivation in physical education. *Journal of Sport & Exercise Psychology*, 25, 145–160.
- Simons, J., Dewitte, S., & Lens, W. (2004). The role of different types of instrumentality in motivation, study strategies, and performance: Know why you learn, so you'll know what you learn! *British Journal of Educational Psychology*, 74, 343–360.
- Tabachnick, S. E. (2005). The impact of future goals on students' proximal subgoals and on their perceptions of task instrumentality (Doctoral dissertation, University of Oklahoma, 2005). *Dissertation Abstracts International*, 66, 493.
- Ullman, J. B. (1996). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (pp. 709–812). NY: HarperCollins.
- United States Department of Education. (1983). *A nation at risk: The imperative for educational reform*. Retrieved June 22, 2003, from <http://www.ed.gov/pubs/NatAtRisk/risk.html>
- United States Department of Education. (1994). *Goals 2000: Educate America Act*. Retrieved June 22, 2003, from <http://www.ed.gov/legislation/GOAL2000/TheAct/>
- United States Department of Education. (2001). *No Child Left Behind Act*. Retrieved June 22, 2003, from <http://nclb.gov/next/overview/index.html>
- Vansteenkiste, M., Lens, W., & Deci, E. L. (2006). Intrinsic versus extrinsic goal contents in self-determination theory: Another look at the quality of academic motivation. *Educational Psychologist*, 41, 19–31.
- Vansteenkiste, M., Simons, J., Lens, W., Sheldon, K. M., & Deci, E. L. (2004). Motivating learning, performance, and persistence: The synergistic effects of intrinsic goal contents and autonomy-supportive contexts. *Journal of Personality and Social Psychology*, 87, 246–260.
- Vansteenkiste, M., Simons, J., Lens, W., Soenens, B., Matos, L., & Lacante, M. (2004). Less is sometimes more: Goal content matters. *Journal of Educational Psychology*, 96, 755–764.
- Vansteenkiste, M., Simons, J., Soenens, B., & Lens, W. (2004). How to become a persevering exerciser: The importance of providing a clear future goal in an autonomy-supportive way. *Journal of Sport & Exercise Psychology*, 26, 232–249.
- Wenger, E. C., & Snyder, W. M. (2000). Communities of practice: The organizational frontier. *Harvard Business Review*, 78, 139–145.
- Wigfield, A., & Eccles, J. S. (2001). The development of competence beliefs, expectations for success, and achievement valued from childhood through adolescence. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 91–120). San Diego, CA: Academic Press.

Received September 19, 2006

Revision received October 25, 2007

Accepted October 25, 2007 ■

## Addressees of Performance Goals

Albert Ziegler, Markus Dresel, and Heidrun Stoeger  
University of Ulm

As performance goals aim to both procure acknowledgment of one's abilities and to avoid revealing a lack of one's abilities, the authors hypothesized that students hold specific performance goals for different addressees and that there are specific correlational patterns with other motivational constructs. They analyzed a data set of 2,675 pupils (1,248 boys and 1,426 girls) attending Grades 8 and 9 (mean age = 15.0,  $SD = 0.97$ ). The students completed a questionnaire consisting of 12 items measuring performance approach goals and 12 items measuring performance avoidance goals. In each subset, 4 groups of addressees were differentiated: parents, teachers, peers, and the acting individual him/herself. Additionally, several external criteria were measured. The authors concurrently tested theory-driven, structural equation models. Incorporating all 24 items, the best-fitting model was a multitrait-multimethod model, which posited 2 factors for approach and avoidance goals and 4 addressee factors. While performance goals addressing parents showed relationships to maladaptive motivational and learning patterns, performance goals addressing classmates and self showed relationships to adaptive motivational and learning patterns. The relationships between performance goals addressing teachers and external criteria were rather weak and unsystematic.

**Keywords:** motivation, achievement goals, addressees of performance goals, approach and avoidance performance goals

There is a general consensus that in scholastic contexts, achievement goals have a decisive influence on achievement behavior (cf. Pintrich & Schunk, 2002). In the research literature, two distinctions have been made to categorize the different goals students hold (e.g., Elliot & McGregor, 2001; Pintrich, 2000). The primary distinction differentiates between mastery goals (sometimes referred to as *learning goals*) and performance goals (Ames, 1992; Dweck, 1986; Maehr & Midgley, 1991). Students who focus on making progress in learning and developing their knowledge, competences, or abilities seek mastery goals. Students who aim to accumulate acknowledgment of their performances or attempt to avoid having others perceive their lack of abilities are committed to performance goals. Empirical studies indicate that mastery goals and performance goals have different types of effects on learning and achievement behavior. For instance, differential correlations with academic self-concept (Skaalvik, 1997; Vrugt, Oort, & Zeeberg, 2002), effort expenditure (Meece, Blumenfeld, & Hoyle, 1988; Wolters, 1998), depth of learning processes (Ames & Archer, 1988; Meece et al., 1998; Wolters, 2004), task value (Bong,

2001), text anxiety (Linnenbrink, 2005; Middleton & Midgley, 1997), and achievement (Elliot & McGregor, 2001; Harackiewicz, Barron, Tauer, Carter, & Elliot, 2000) could be confirmed. Whereas mastery goals have proven to be adaptive in just about all studies, the findings on performance goals were somewhat less distinct. Some studies showed negative relationships between performance goals and the aspects of student learning mentioned above; other empirical studies found no relationships or even positive relationships (see Elliot, 1999, for an overview). In order to explain these ambiguities associated with performance goals, a second distinction was introduced. It considers the positive versus negative valence of the state or situation the goal setting process focuses on (Elliot, 1999; Elliot & Harackiewicz, 1996; Elliot & Sheldon, 1997; Middleton & Midgley, 1997). This distinction differentiates between an approach component—in the sense of an approach toward desirable states or situations—and an avoidance component—in the sense of an avoidance of undesirable states or situations. Within the past decade, this distinction has been established for performance goals, whereby it could be shown that approach goals and avoidance goals are positively correlated, but nonetheless separate components (Elliot & Harackiewicz, 1996; Elliot & Sheldon, 1997; Middleton & Midgley, 1997). Research on the consequences of performance approach goals and performance avoidance goals showed that the above-described negative effects of performance goals can primarily be ascribed to the avoidance component. It is negatively related to scholastic performance, academic self-concept, task value, and effort expenditure as well as positively related to test anxiety, maladaptive attributional style, and surface processing (e.g., Dresel, 2001; Elliot & Church, 1997; Elliot & McGregor, 2001; Elliot & Sheldon, 1997; Middleton & Midgley, 1997; Pintrich, 2000). The findings pertaining to the approach component were, in contrast, less consistent (for an

---

Albert Ziegler, Markus Dresel, and Heidrun Stoeger, Department of Educational Psychology, University of Ulm, Ulm, Germany.

Markus Dresel is now at the Department of Psychology, University of Augsburg, Augsburg, Germany. Heidrun Stoeger is now at the Department of School Research, University of Regensburg, Regensburg, Germany.

An earlier version of this article was presented at the 11th Conference of the European Association for Research on Learning and Instruction, Nicosia, Cyprus, August 2005. We thank Thomas D. Raul for his assistance in the translation of this article.

Correspondence concerning this article should be addressed to Albert Ziegler, Department of Educational Psychology, University of Ulm, Albert-Einstein-Allee 47, 89069 Ulm, Germany. E-mail: albert.ziegler@uni-ulm.de



overview, see Midgley, Kaplan, & Middleton, 2001). Although positive correlations could be confirmed with achievement and academic self-concept, findings regarding task value, effort expenditure, and depth of learning processes varied between null and moderately positive correlations (Harackiewicz, Barron, Pintrich, Elliot, & Thrash, 2002; Kaplan & Middleton, 2002; Midgley et al., 2001).

Recently, researchers also made a distinction between an approach and an avoidance component within mastery goals, resulting in a full  $2 \times 2$  framework of achievement goals (Elliot & McGregor, 2001; Pintrich, 2000). It can be argued that students may also focus on the undesirable state of misunderstanding and not mastering a task and therefore adapt their learning in accordance with avoidance goals. Research has provided some evidence that mastery avoidance goals are associated with a more negative motivational set than mastery approach goals and a more positive motivational set than performance avoidance goals (Elliot & McGregor, 2001).

In the present work, we focus on performance approach and performance avoidance goals. More specifically, our aim was to test whether the structures and the relationships of performance goals depend on the various social referents to which they are addressed. Our work is anchored in the two-component definition of performance goals advanced by Elliot (1999)—namely social comparison and appearance—and we are particularly interested in the significance of the appearance component. As stated above, performance goals focus on the demonstration of competence or, in the case of the avoidance component, the attempt to avoid having others perceive a lack of competence. One can presume that an important role is played by who these “others” are. Furthermore, we hypothesized that addressee-specific performance approach and avoidance goals may be related to other constructs in a specific manner, reflecting the relevance of the specific social instance on which the current goal is focused. We hold the view that a systematic consideration of the various addressees of performance approach goals and performance avoidance goals may lead to a deeper understanding of the goals that students hold in the social context of scholastic learning as well as the consequences of goal setting processes.

### Addressee-Specific (Interpersonal) Performance Goals

In the context of scholastic learning, there are three major groups of important others: teachers, classmates, and parents. Teachers define most of the learning tasks, provide help, bring about formal and informal performance situations (such as tests or oral classroom questions), define the major standards for evaluation of performances, provide students with verbal (public, face-to-face) or written feedback, and often react to students' performances with different emotions. In other words, they are typically the most salient persons in the classroom setting. Classmates are typically copresent in all performance situations but are absent in face-to-face conversations with the teacher. They define a second bundle of standards (e.g., “being good in school is uncool”) and reward or punish students with verbal statements and with emotions such as sympathy or antipathy. Parents are commonly absent in all classroom learning and performance situations. However, they provide students with a learning environment at home, occasionally undertake learning activities together with their child,

codefine standards for the evaluation of school performances, often reward good performances and punish bad performances, and react in emotional terms. To summarize this list of characteristics (for an overview, see Wentzel, 1998, 1999), these groups of important others share different situations with the student, provide different standards of performance evaluation, bring up rewards and punishments based on presumably different standards, and place different expectations on the student.

Based on the assumption that students are aware of these differing standards, reward systems, and expectations, it is plausible that the salience of one's competence or competence deficits are weighted differently for different groups of important others. As the aim to either have one's competences acknowledged or avoid having others perceive a lack of one's competences builds the core concept of performance goals, this assumption implies that students probably differ in their performance goal setting process among different groups of important others. In other words, different performance goals may be directed to different addressees. For instance, a student may be concerned about the perceptions of his or her parents, but he or she may not care about the perceptions of his or her classmates.

The significance of addressees for learning and the goal setting process has also been investigated by other authors and has resulted in some degree of empirical validation (Harris, 1995; Wentzel, 1998, 1999); nevertheless, to this point a systematic evaluation has not yet been attempted. This is also reflected in the measuring instruments that have been used to assess performance goals. Although some measuring instruments explicitly include teachers or pupils as addressees of learning and achievement behavior (e.g., Midgley et al., 1996; Miller, Greene, Montalvo, Ravindran, & Nichols, 1996), no systematic examination on the influence of the individual addressees was ever undertaken. In some cases, only classmates (e.g., DeBacker Roedel, Schraw, & Plake, 1994) or teachers and classmates (e.g., Midgley et al., 1996) were included in the items as addressees, but not parents.

Empirical indications for addressee-specific goals have been supplied by Urda and Mestas (2006) with their qualitative study. In structured interviews, they questioned pupils about their individual reasons for pursuing performance approach goals (e.g., “I want to do better than other students in this class”) and performance avoidance goals (e.g., “It is important to me not to do worse than other students in this class”). An analysis of the answers disclosed a variety of purposes behind students' goal pursuits. Included here were purposes that were directed toward specific addressees, such as wanting to please parents or to silence naysaying peers. These were categorized by the authors as interpersonal performance goals (see also Wentzel, 1998, 1999). Additional support for addressee-specific goal setting processes are inherent in findings that show that individual goals depend on classroom goal structures, general instructional formats, parents' goals, and peer group standards (e.g., Church, Elliot, & Gable, 2001; Grusec & Goodnow, 1994; Kinderman, 1993).

### Intrapersonal Performance Goals: The Individual Him/ Herself as an Addressee

In addition to important others, we assume that the acting individual may him/herself be an addressee of performance goals. In our opinion, intrapersonal or self-addressed performance goals



are goals that imply a normative standard of competence definition (resulting from the performances of others), but—in contrast to performance goals that are directed to external addressees—they do not associate a positive appearance to others with a desirable state or a negative appearance to others with an undesirable state. Examples for intrapersonal performance goals could be wanting to be satisfied with oneself for attaining good grades, wanting to be happy about one's performances, or wanting to avoid dissatisfaction with bad performances (see Urdan & Mestas, 2006, for a similar conceptualization of intrapersonal performance goals and preliminary qualitative evidence for their existence).

In contrast to interpersonal performance goals, which are assumed to trigger the use of techniques to control the appearance component (such as selective communication or the utilization of excuses), intrapersonal performance goals may result in the application of self-enhancement processes with self-confidence as an anticipated state. The anticipated state is also the decisive characteristic that differentiates intrapersonal performance goals from mastery goals: While the former focus on the consummate magnitude of one's competence and its appreciation, the latter focus on the growth or development of one's competences (see Dweck, 1999). As a consequence, intrapersonal or self-addressed performance goals should act as a guide in the procurement of information to protect self-confidence and self-worth, while mastery goals should act as a guide in the search for realistic information to enable optimal learning (see Dweck & Elliott, 1983). Nevertheless, we expect that self-addressed performance goals are more closely related to mastery goals than externally addressed performance goals. This is because self-addressed performance goals, plausibly, not only underlie a normative standard but also an absolute standard of competence evaluation (resulting from the task requirements), similar to mastery goals for which an intrapersonal standard (resulting from one's past performances) and/or an absolute standard is applied (see Ames, 1992; Dweck, 1999; Maehr, 1989; Elliot & McGregor, 2001; cf. Pintrich & Schunk, 2002).

### Research Questions and Hypotheses

The main purpose of the present study was to test whether performance approach and avoidance goals can be specified for four addressee groups (parents, teachers, classmates, and the student him/herself) in the context of scholastic learning. To validate this distinction, we analyzed the associations between addressee-specific performance goals and other motivational and emotional facets of the learning process, namely mastery goals, academic self-concept, achievement, effort expenditure, depth of learning processes, task value, and test anxiety.

Four major hypotheses were tested: (1) Performance goals are differentiated for different addressee groups within both performance approach goals and performance avoidance goals. (2) Performance goals are divided in performance approach goals and performance avoidance goals within each addressee group. (3) The simultaneous incorporation of the distinction between an approach and an avoidance component as well as the distinction among different addressees results in a better representation of performance goals than the exclusive incorporation of only one of the two distinctions. (4) Different addressee-specific goals show different relationships to mastery goals and other aspects of the learning process.

## Method

### Participants

Analyses of the present study are based on a sample that is part of the calibration sample for a new measuring instrument, the Ulm Motivational Test Battery (Ziegler, Dresel, Schober, & Stoeger, 2005). Our analyses included all eighth graders and ninth graders in the calibration sample for which a complete data set was amassed (1,354 eighth graders and 1,321 ninth graders). The 2,675 students were attending one of the three major types of German public schools: Hauptschule (lowest achievement level; 34.7%), Realschule (average achievement level; 33.5%) and Gymnasium (highest achievement level; 31.8%).<sup>1</sup> Students participated voluntarily in the investigation and had obtained permission from their parents. The investigation was conducted as a paper-and-pencil test during regular classroom instruction, and took about 45 min to complete. The mean age of the students came to 15.0 years ( $SD = 0.97$ ), 53.3% of whom were girls.

### Measuring Instruments

**Performance goals.** To assess the performance approach goals and performance avoidance goals specifically being addressed to various members of the learning environment or the acting individual him/herself, we used a previously developed questionnaire. Research (Schober, Ziegler, & Dresel, 2001; see also Ziegler & Stoeger, 2002) confirmed that this measuring instrument is consistent with prevailing quality criteria such as internal consistency and criterion validity. The 24-item questionnaire systematically crosses the two performance goal components (approach and avoidance component), with the four major addressee groups relevant in the scholastic context. These addressee groups are the parents of the students, their teachers, their classmates, and the acting individual him/herself. Crossing the two components with the four addressee group results in eight combinations. For each of these combinations, the questionnaire contains three items that focus on different anticipated reactions of addressees (e.g., taking notice of the competences of the student, evaluating these competences, reacting in an emotional manner). All items were measured with a 6-point Likert-type scale, ranging from 1 (*I disagree completely*) to 6 (*I agree completely*). All 24 items are listed in Table 1, accompanied by their respective goal component and addressee group.

**External criteria.** To examine whether addressee-specific performance goals are associated with other constructs in a specific manner, we measured seven external criteria that are known to be related to performance goals. Unless specified otherwise in the following, all items on the respective measurements were presented along a 6-point Likert-type scale, ranging from 1 (*I disagree completely*) to 6 (*I agree completely*). To assess mastery goals, we

<sup>1</sup> In the German public school system following the fourth grade, pupils are allocated to one of the three school types named above on the basis of academic achievement. The curriculum in a Hauptschule lasts for a period of 5 years, the curriculum in a Realschule lasts 6 years, and in a Gymnasium—depending on federal state—students study for 8 or 9 years. The first two school types train students for future occupations (apprenticeship); the third serves as preparation for university studies.



Table 1  
*Descriptive Statistics and Item Texts for the 24 Items Assessing Performance Goals*

Addressee and item	Item text ("In school ...")	<i>M</i>	<i>SD</i>
Approach component			
Parents			
Ap-P1	... I want my parents to notice how good I am.	4.13	1.23
Ap-P2	... I want my parents to be proud of me because I am good.	4.67	1.16
Ap-P3	... I want my parents to praise me because I am good.	4.51	1.18
Teacher			
Ap-T1	... I want my teacher to notice how good I am.	4.18	1.26
Ap-T2	... I want my teacher to like me because I am good.	3.24	1.42
Ap-T3	... I want my teacher to praise me because I am good.	3.87	1.33
Classmates			
Ap-C1	... I want my classmates to notice how good I am.	3.35	1.31
Ap-C2	... I want my classmates to like me because I am good.	2.96	1.38
Ap-C3	... I want my classmates to admire me because I am good.	2.83	1.37
Self			
Ap-S1	... I want to get a good grade.	5.07	0.95
Ap-S2	... I want to be able to be happy about a good grade.	5.39	0.80
Ap-S3	... I want to be satisfied with myself, because I got a good grade.	5.35	0.83
Avoidance component			
Parents			
Av-P1	... I do not want my parents to notice that that I can't do something.	2.82	1.35
Av-P2	... I want to avoid disappointing my parents because I am bad.	4.20	1.39
Av-P3	... I do not want my parents to reproach me because I am bad.	4.23	1.53
Teacher			
Av-T1	... I do not want my teacher to notice that I can't do something.	3.04	1.36
Av-T2	... I want to avoid having my teacher not like me because I am bad.	3.31	1.58
Av-T3	... I do not want my teacher to reproach me because I am bad.	4.06	1.58
Classmates			
Av-C1	... I do not want my classmates to notice that I can't do something.	2.86	1.27
Av-C2	... I want to avoid having my classmates not like me anymore because I am bad.	3.26	1.67
Av-C3	... I do not want my classmates to laugh at me because I am bad.	3.87	1.70
Self			
Av-S1	... I do not want to get a bad grade.	4.02	1.39
Av-S2	... I do not want to be sad about a bad grade.	4.21	1.32
Av-S3	... I want to avoid being dissatisfied with myself for getting a bad grade.	5.18	0.97

Note. *N* = 2,675. Scale range for all items: 1–6.

used the 6-item scale that is also included in the questionnaire for the assessment of the goals stated above (Schober et al., 2001). In this scale, mastery goals are exclusively operationalized as approach goals (sample item: "In school I want to learn as much as possible"). Academic self-concept was assessed with the scale "Confidence in one's own competence" (Dweck & Henderson, 1988). This scale consists of four item pairs containing two statements corresponding to a positive self-evaluation and a negative self-evaluation. The two poles of a 6-point answer scale are formulated as statements (e.g., "I am not sure that I am good enough to be successful in school" and "I am sure that I am good enough to be successful in school"). Achievement was operationalized by averaging report card grades obtained in the subjects German (native language), English (foreign language), and mathematics on the previous year's report cards. The German grading scale ranges from 1 (very good) to 6 (unsatisfactory). The resulting scale was recoded, so that a higher score represented better achievement. Effort expenditure was measured with a 12-item scale (Ziegler et al., 2005). The scale offers insight into the amount of effort students apply to their learning (sample item: "I spend a lot of time at home doing school exercises"). Depth of learning processes was operationalized with a scale developed by Gold and Souvignier

(2004) and was assessed with seven items (sample item: "While studying I try to find examples to match the material"). Task value was measured with a scale consisting of three items that were drawn from Ziegler et al. (2005). This scale primarily measures the utility of learning (sample item: "The things you learn in school are useful later on"). Test anxiety was measured with a six-item scale, which had been evaluated in earlier studies (Schober et al., 2001; Ziegler et al., 2005). Similar to the performance goal measures, test anxiety items were associated with specific addressees (sample item: "When I think about school, I am afraid that my teacher will notice that I can't do something").

### Analyses

To test several concurrent hypotheses concerning the differentiation of performance goals among addressee groups and their relation to external criteria, we performed confirmatory factor analyses using LISREL 8.51 (Jöreskog & Sörbom, 2001). The analyses were based on covariance matrices and used maximum-likelihood estimation. In each set of analyses, several theory driven models were compared against one another using chi-square tests

Table 2  
*Item Consistencies, Descriptive Statistics, and Zero-Order Correlations for the External Criteria*

External criterion	$\alpha$	$M$	$SD$	1	2	3	4	5	6	7
1. Mastery goals	.80	4.52	0.80	—						
2. Academic self-concept	.74	4.12	0.92	.26	—					
3. Achievement	.69	4.10	0.68	.12	.41	—				
4. Effort expenditure	.90	3.74	0.82	.55	.21	.21	—			
5. Depth of learning processes	.86	3.23	0.96	.37	.21	.07	.36	—		
6. Task value	.82	4.29	0.90	.59	.22	.10	.50	.31	—	
7. Test anxiety	.86	2.60	1.00	.00	-.32	-.19	-.01	.10	-.05	—

Note.  $N = 2,675$ . All  $r_s > .06$ ,  $p_s < .001$ .

(e.g., Byrne, 1998). Hypothesized and alternative models are described in the Results section.

## Results

### *Descriptive Statistics*

Table 1 contains means and standard deviations for all 24 performance goal items. Scale consistencies, descriptive statistics, and zero-order correlations for the external criteria examined are presented in Table 2.

### *Model Tests Regarding Approach and Avoidance Components*

The first confirmatory factor analysis set was performed to test whether a differentiation among the four addressee groups could be confirmed in the performance approach component, in the performance avoidance component, or in both (Hypothesis 1). We conducted two separate sets of analyses, each including the 12 respective, component specific items (see Table 1). Figure 1 depicts the three models that were tested for each component: (a) total rating factor, in which all 12 items load on one latent variable; (b) others-self, in which the three self-addressed items load on one latent variable and the remaining nine items that address others load on a second latent variable; (c) four addressee groups, in which the parents, teacher, classmates, and self-items load on their respective latent variables, resulting in four correlated, addressee-specific factors. This is the hypothesized model.

As displayed in Table 3, the results indicate clear support for a distinction among several addressees regarding performance approach goals. In contrast to the two alternative models, the hypothesized four addressee groups' model revealed a good fit to the performance approach item data.<sup>2</sup> Model comparisons indicated that the hypothesized model provided a better fit than the other two models. Moderate to large factor correlations were observed among the three latent factors concerning other persons (see Table 4). For self-addressed performance approach goals, we found small to moderate correlations with externally addressed performance approach goals, whereby the highest correlation was observed with the parent factor.

Results concerning the performance avoidance component revealed a somewhat different pattern: Although model comparisons disclosed significant advantages for the hypothesized four addressee groups model, the model fit differences among the three models were smaller than those found for performance approach

goals. Additionally, none of the three models revealed an adequate fit to the data. Finally, correlations calculated for the latent factors of the hypothesized model were generally larger for performance avoidance goals than those for performance approach goals. In one case, a correlation greater than 1 was even calculated, resulting in a latent factor correlation matrix, which was not positive definite (see Table 4).

Inspection of the modification indices of the performance avoidance models revealed strong evidence for substantial correlated errors between two item pairs (modification indices  $> 200$ ), which focus on identical anticipated reactions of addressees (Items Av-P3 and Av-T3 are both focusing on reproaching; Items Av-P1 and Av-T1 are both focusing on taking notice of deficient competences; see Table 1). Within an exploratory framework, we repeated the analyses with setting the two error correlations free. Significant correlated errors for the two item pairs were estimated for all three models ( $\delta_s = .22$ – $.24$  and  $\delta_s = .25$ – $.26$ , respectively). Results indicated an acceptable model fit for the four addressee groups model and slightly worse model fits for the alternative models (see Table 3). The hypothesized differences between the models respecting data fit remained significant. The latent factor correlation matrix of the four addressee group model was now positive definite and consisted of slightly lower values than in the model without correlated errors (see Table 4). Nevertheless, latent correlations among the four addressee factors respecting avoidance goals were still substantially larger than those respecting performance approach goals.

### *Model Tests Regarding Addressee Groups*

To additionally test whether the approach-avoidance distinction is also valid for specific addressees (Hypothesis 2), we conducted modeling within the four addressee groups. For each addressee group, we tested two models with the respective three approach and three avoidance items: total rating factor, in which all six items load on one latent variable, and approach-avoidance, in which the approach items load on one latent variable and the avoidance items load on another. This is the hypothesized model.

For parents, teachers, and classmates as addressees of performance goals, results clearly indicate that a distinction between approach and avoidance goals holds true. For all three groups, the

<sup>2</sup> Because the sample size is very large, in this and all subsequent analyses significant chi-square tests are not unexpected and should not be interpreted as an indicator of bad model fit (see Byrne, 1998).



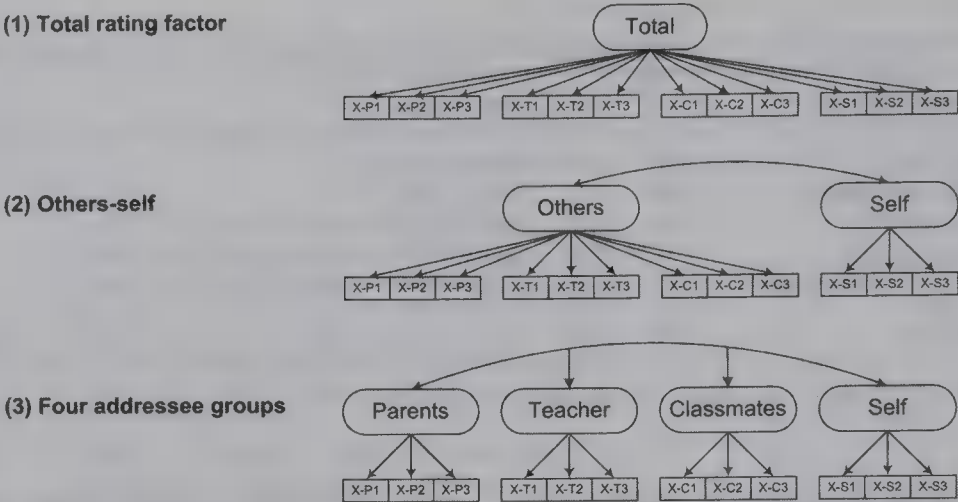


Figure 1. Hypothesized models for both performance approach and performance avoidance components. For reasons of clarity, uniquenesses are not included in the figure. First set of analyses (1): X = performance approach item. Second set of analyses (2): X = performance avoidance item. P = parents item; T = teacher item; C = classmates item; S = self-item.

approach–avoidance model revealed sufficient fit indices for the most part (parents: comparative fit index [CFI] = .99, Tucker–Lewis index [TLI] = .99; teacher: CFI = .92, TLI = .85; classmates: CFI = .94, TLI = .89), whereas the total rating factor model resulted in lower fit indices (parents: CFI = .94, TLI = .91; teacher: CFI = .89, TLI = .82; classmates: CFI = .86, TLI = .76). Model comparisons revealed for all three addressee groups that the two component model fits the sample data significantly better than the model featuring a total rating factor,  $\chi^2(1) > 136.7, ps < .001$ . As expected, the modeling revealed moderate to high latent correlations between performance approach goals and performance avoidance goals. The correlations between the two latent factors were  $\varphi = .70$  for parent-related goals,  $\varphi = .79$  for teacher-related goals, and  $\varphi = .66$  for goals that are directed toward classmates. In contrast to these three external addressee groups, results regarding goals that are directed to the acting person him/herself are somewhat different: Both hypothesized models fit the data in a comparable manner (total rating factor model: CFI = .94, TLI = .90; approach–avoidance model: CFI = .94, TLI = .89). The differences between the fit indices were not significant,  $\chi^2(1) =$

0.4,  $p = .52$ . Moreover, the latent correlation between the two components in the approach–avoidance model was very high ( $\varphi = .92$ ). Obviously, the approach–avoidance distinction is less evident for self-addressed performance goals than for externally addressed performance goals.

Model Tests Incorporating All Facets

In the next step, we tested a series of models that incorporated all 24 items (Hypothesis 3; see Figure 2): approach–avoidance, in which the 12 approach items load on one latent variable and the 12 avoidance items load on another; four addressee groups, in which the parents, teacher, classmates, and self-items (six items in each case) load on their respective latent variables, resulting in four correlated addressee-specific factors; approach–avoidance by four addressee groups, in which eight component and addressee-specific factors were defined, each having three items; and approach–avoidance and four addressee groups, in which the 12 approach items and the 12 avoidance items load on their respective latent variables and additionally in which the parents, teacher,

Table 3  
Results From Confirmatory Factor Analyses for Approach Goals and for Avoidance Goals

Model	<i>df</i> or $\Delta df$	Approach goals				Avoidance goals			
		$\chi^2$ or $\Delta\chi^2$	RMSEA	CFI	TLI	$\chi^2$ or $\Delta\chi^2$	RMSEA	CFI	TLI
Model fit									
1. Total rating factor	54 (52)	4656.3*	.18	.74	.69	1237.1* (688.2*)	.09 (.07)	.84 (.91)	.80 (.88)
2. Others–self	53 (51)	3046.3*	.15	.83	.79	1197.7* (646.7*)	.09 (.07)	.84 (.91)	.80 (.89)
3. Four addressee groups	48 (46)	815.5*	.08	.95	.93	1102.1* (539.0*)	.09 (.06)	.86 (.93)	.80 (.90)
Model comparison									
Model 3 vs. Model 1	6 (6)	3840.8*				135.0* (149.2*)			
Model 3 vs. Model 2	5 (5)	2230.8*				95.7* (107.7*)			

Note.  $N = 2,675$ . Analyses for avoidance goals were repeated with setting free the correlated errors between two item pairs each focusing identical reactions of addresses. Results are presented in parentheses. RMSEA = root-mean-square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index.  
\*  $p < .001$ .

Table 4  
*Latent Factor Correlations in the Four Addressee Groups  
 Models for Approach Goals (Above the Diagonal) and for  
 Avoidance Goals (Below the Diagonal)*

Addressee of goals	1	2	3	4
1. Parents	—	.82	.57	.62
2. Teacher	1.03 (.86)	—	.81	.45
3. Classmates	.82 (.79)	.96 (.94)	—	.25
4. Self	.77 (.75)	.75 (.75)	.64 (.65)	—

*Note.*  $N = 2,675$ . Analyses for avoidance goals were repeated with setting free the correlated errors between two item pairs each focusing identical reactions of addressees. Results are presented in parentheses.

classmates, and self-items load on their respective four addressee factors. This is the multitrait-multimethod model (MTMM model), which posits two loadings for each item on a component factor and on an addressee factor (refer to Byrne, 1998).

The fit indices for the four models are presented in Table 5. In line with our expectations, models that exclusively posit approach and avoidance components or the four addressee groups do not fit the data nearly as well as models that incorporate the influences of both performance goal components and addressee groups. The model crossing approach and avoidance components with addressee groups (resulting in eight latent variables) fits substantially better, but the model with the best fit is the MTMM model, which simultaneously posits two component factors and four addressee factors. CFI and TLI values above .90 also indicate that the approach-avoidance and four addressee groups model is the only model with an adequate fit to the data.<sup>3</sup>

Latent factor correlations for the MTMM model are displayed in Table 6. These correlations specify the relationship between the approach and avoidance components controlled for several addressee groups and vice versa. Results reveal that the approach component and the avoidance component correlate high positive, when the addressee groups of performance goals are controlled for. On the other hand, moderate to high latent correlations between performance goals addressing differential social referents emerged when the distinction between approach goals and avoidance goals was controlled for. The latent correlation between parent-addressed goals and teacher-addressed goals was particularly high. However, latent correlations incorporating the acting individual him/herself as the addressee of performance goals were substantially lower, especially for associations with teacher- and classmate-directed goals.

### *Relationships With External Criteria*

To examine relationships between addressee-specific performance goals and the external criteria observed (Hypothesis 4), we expanded the model with the best fit (approach-avoidance and four addressee groups) to a regression model with free regression paths from both component factors as well as from each addressee group factor to one latent endogenous variable. We ran the analysis separately for each of the seven external criteria as an endogenous variable. Central to our hypotheses are the regression coefficients of the structural model, which are displayed in Table 7.

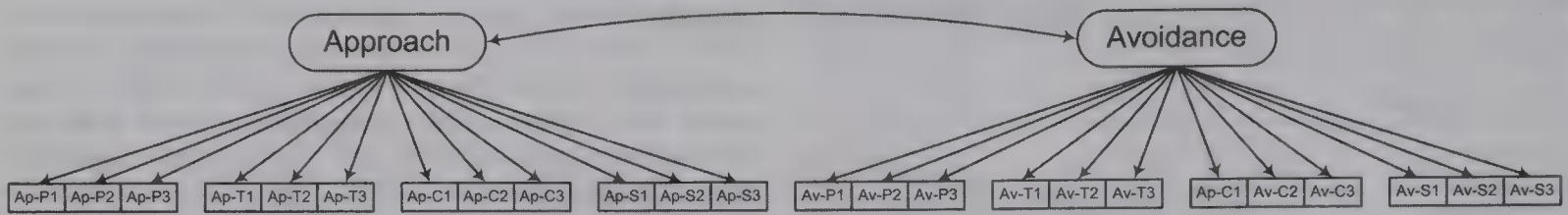
While systematic influences of several addressees were controlled in these models, the prediction of the seven external criteria from performance approach goals and performance avoidance goals provided highly valid estimations of associations with other motivational variables. Here, a distinct pattern occurred for approach and avoidance goals. Performance approach goals predicted mastery goals, academic self-concept, effort expenditure, depth of learning processes, and task value in a positive sense. In contrast, performance avoidance goals demonstrated weak to moderately negative relationships with mastery goals, academic self-concept, achievement, effort expenditure, and task value. Additionally, performance avoidance goals showed a clearly positive relationship to test anxiety, while performance approach goals could not predict this significantly.

The regression coefficients of the four factors corresponding to the four addressee groups are relevant for our hypothesis that addressee specific relationships with external criteria do exist. Since the model structure incorporates a control for approach and avoidance components, these are estimates of the unique effects of each of the four addressee groups, in that they partial out the above described global effects of the two goal components. As hypothesized, results indicate different relationships, contingent on the addressees of performance goals for all external criteria. Of particular interest for the achievement goal literature may be the regression coefficients associated with mastery goals. Here, our results reveal a negative association with parent-addressed performance goals and positive associations with classmate-addressed and self-addressed performance goals, whereby the latter was especially high. Similar relationships occurred for task value—a supplemental negative effect was, admittedly, observed for teacher-directed performance goals. With respect to academic self-concept and achievement, self-addressed performance goals again served as a significant positive predictor. Academic self-concept was additionally predicted by parent-directed performance goals in a negative sense. While predicting effort expenditure, goals that were addressed to classmates or to the student him/herself turned out to be positive predictors, while regression coefficients for parent- and teacher-addressed goals were insignificant. The regression of depth of learning processes again displayed a differential importance for several of the addressees: While parent- and classmate-addressed goals did not predict the extent of deep learning processes, teacher- and self-addressed goals predicted it positively. Finally, a differential pattern also appeared for test anxiety: It was positively associated with performance goals, which are directed toward parents, and negatively associated with performance goals aimed at one's self. While the respective regression coefficient was relatively low for the individual him/herself as an addressee of performance goals, a substantially higher coefficient was observed for parents as addressees of performance goals.

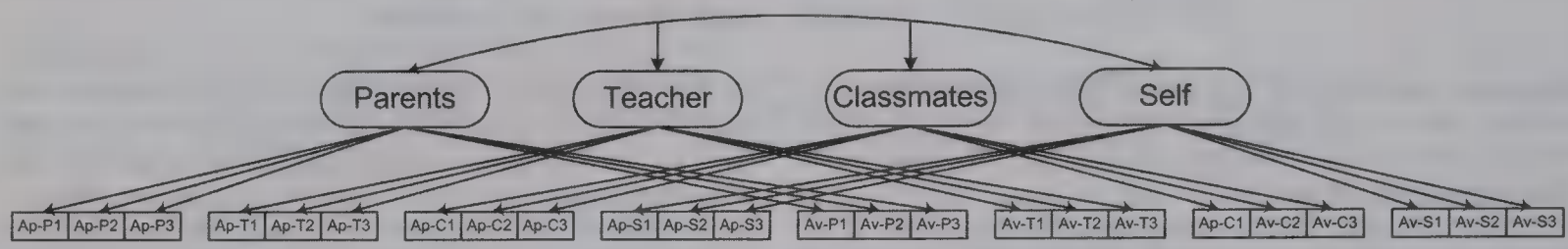
<sup>3</sup> Analogous to the models in the avoidance component, large modification indices respecting correlated errors between the above named items were observed (modification indices > 126.7). Since the fit of the MTMM model was satisfactory before the correlated uniquenesses was set free, and the fit of the remaining models would not improve substantially through such exploratory framed specifications (CFIs < .89, TLIs < .88), we did not include them in the models.



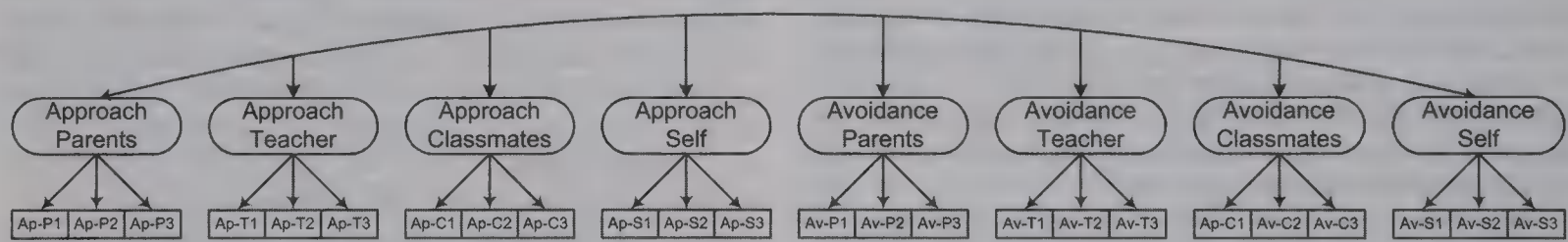
(1) Approach-avoidance



(2) Four addressee groups



(3) Approach-avoidance by four addressee groups



(4) Approach-avoidance and four addressee groups

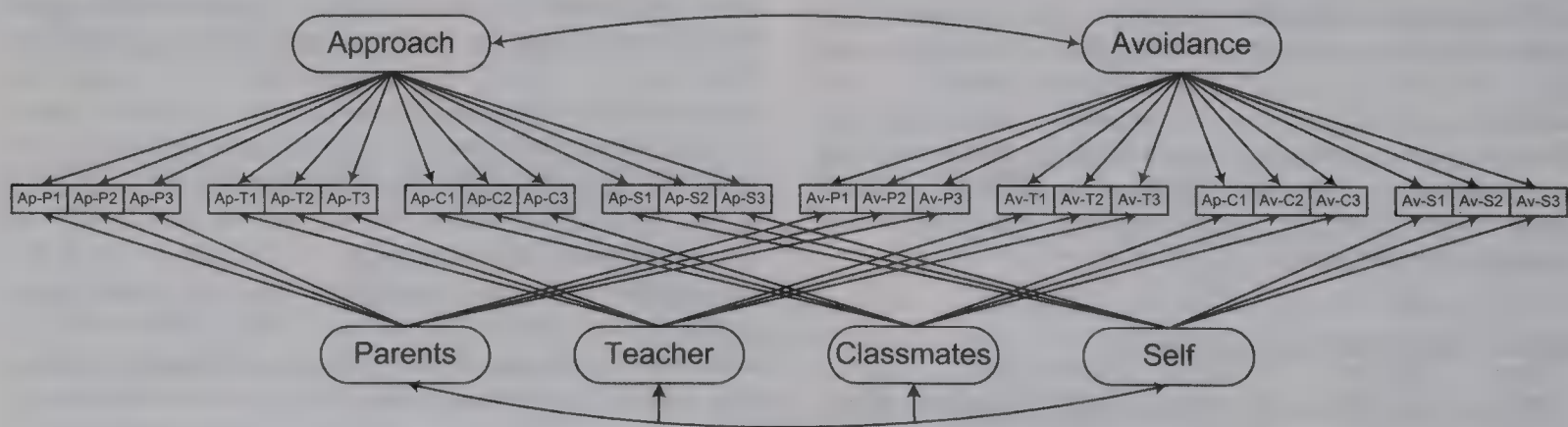


Figure 2. Hypothesized models incorporating all facets of performance goals. For reasons of clarity, uniquenesses are not included in the figure. Ap = approach; Av = avoidance; P = parents; S = self; T = teacher; C = classmates.

Table 5  
*Model Fit and Model Comparison Results From Confirmatory Factor Analyses With all Facets of Performance Goals*

Model	df or $\Delta df$	$\chi^2$ or $\Delta\chi^2$	RMSEA	CFI	TLI
Model fit					
1. Approach-avoidance	251	9159.1*	.12	.73	.71
2. Four addressee groups	246	6532.5*	.10	.80	.77
3. Approach-avoidance by four addressee groups	224	3427.1*	.07	.88	.86
4. Approach-avoidance and four addressee groups	221	2011.4*	.06	.93	.91
Model comparison					
Model 3 vs. Model 1	27	5732.0*			
Model 3 vs. Model 2	22	3105.4*			
Model 4 vs. Model 1	30	7147.7*			
Model 4 vs. Model 2	25	4521.1*			
Model 4 vs. Model 3	3	1415.7*			

Note.  $N = 2,675$ . RMSEA = root-mean-square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis index.

\*  $p < .001$ .

## Discussion

The work at hand pursued the central idea that performance goals could be addressee specific. This idea is based on the differing situations, standards, reward systems, expectations, and emotional reactions relating to various groups of important others, whereby in the context of scholastic learning, parents, teachers, and classmates are of prime relevance (Wentzel, 1998, 1999). Moreover, we assumed that the individual him/herself could be an addressee of performance goals (refer also to Urdan & Mestas, 2006). The study aimed to clarify whether the differentiation among these four addressee groups holds true for performance approach and performance avoidance goals and to test whether an additional and systematic consideration of addressee groups results in a better representation and understanding of performance goal processes (including associations with other motivational and emotional facets of learning) than the exclusive incorporation of the approach-avoidance distinction.

With respect to our first hypothesis, it was shown that different addressee groups comprise separate factors for performance goals. This was clearly valid for performance approach goals and to a somewhat lesser degree also for performance avoidance goals. For both components, a four-addressee-group model offers a significantly better model fit than either a total-rating-factor model or a model that differentiates between interpersonal and intrapersonal

performance goals (see Urdan & Mestas, 2006). One can conclude that pupils differentiate among various addressees while building up performance goals. Thus, they distinguish to whom they want to appear to be competent or not to appear to be incompetent. Moreover, they often hold a separate class of performance goals, which do not target how they appear to other persons, but are exclusively intrapersonal in nature. Noteworthy is that, for performance avoidance goals, the four addressee factors were more closely associated with each other than those for performance approach goals. Moreover, differences between the hypothesized model and the alternative models were less substantive for performance avoidance goals. This could be an indication that the anticipation of positive states (which is inherent in approach goals) is more specifically focused on certain positive outcomes (e.g., the impression given to a specific person that one is competent) than the anticipation of negative states (a constituent of avoidance goals), which apparently is focused more generally on the negative appearance of student competences.

The results concerning our second hypothesis revealed that the distinction between an approach and an avoidance component holds true for all external addressee groups of performance goals. For intrapersonal performance goals, the validity of the approach-avoidance distinction did not hold. One can infer that the approach-avoidance distinction is definitely relevant when students aim to be viewed by others as competent or aim to avoid being viewed as incompetent, but this distinction is less striking when the appearance component is absent, in other words when students pursue the intrapersonal goal of being confident or not being confident about their competences (cf. Elliot, 1999).

Testing models with all facets of performance goals resulted in clear indications that simultaneous incorporation of the distinction between an approach and an avoidance component as well as the distinction among different addressees result in a better fit to sample data than the exclusive application of only one of the two distinctions. According to these results, taking different addressees into account provides researchers with a better understanding of goal setting processes. A particularly good fit resulted for the MTMM model, which independently posits two component factors

Table 6  
*Latent Factor Correlations in the Multitrait-Multimethod Model Approach-Avoidance and Four Addressee Groups*

Variable	1	2	3	4	5	6
1. Approach component	—					
2. Avoidance component	.67	—				
3. Parents			—			
4. Teacher			.76	—		
5. Classmates			.59	.67	—	
6. Self			.64	.48	.38	—

Note.  $N = 2,675$ .



Table 7  
*Standardized Coefficients of the Regression of External Criteria on Approach and Avoidance Components as well as on the Four Addressee Groups*

External criterion	Goal component		Addressees of goals			
	Approach	Avoidance	Parents	Teacher	Classmates	Self
Mastery goals	.34	-.09	-.16	.10 <sup>a</sup>	.26	.55
Academic self-concept	.35	-.38	-.11	.06 <sup>a</sup>	.02 <sup>a</sup>	.19
Achievement	.05 <sup>a</sup>	-.14	-.01 <sup>a</sup>	.01 <sup>a</sup>	.00 <sup>a</sup>	.15
Effort expenditure	.29	-.14	.03 <sup>a</sup>	-.01 <sup>a</sup>	.25	.28
Depth of learning processes	.22	-.07 <sup>a</sup>	-.08 <sup>a</sup>	.15	.03 <sup>a</sup>	.20
Task value	.40	-.11	-.13	-.16	.30	.36
Test anxiety	.02 <sup>a</sup>	.45	.31	.07 <sup>a</sup>	.08 <sup>a</sup>	-.13

Note.  $N = 2675$ . Regression coefficients were estimated by expending the multitrait-multimethod model approach-avoidance and four addressee groups with one external criterion as an endogenous variable.

<sup>a</sup> not significant.

and four addressee factors. These findings imply that addressee specification could be a distinction, similar in importance to the approach-avoidance distinction, while being relatively independent from it. Accordingly, results clearly revealed that the differentiation among various addressees can hardly replace the approach-avoidance distinction in achievement goal theory. Addressee specification could, rather, provide a justifiable supplement and refinement of the existing theoretical framework. For instance, when researchers are interested in associations between performance goals and other motivational and emotional facets of learning, these should be systematically controlled for several addressees. In fact, a contribution to the existing literature on achievement goal theory may be provided by the results compiled on the associations of performance approach goals and performance avoidance goals with external criteria under systematic control for addressee-specific influences. Our addressee-controlled results may help to clarify the contrary association patterns regarding performance approach and performance avoidance goals with various aspects of adaptive and maladaptive learning (for an overview, see Midgley et al., 2001).

With our forth hypotheses, we focused on addressee-specific associations of performance goals with the above designated external criteria. As expected, different addressee groups of performance goals were of different importance with respect to relationships with mastery goals and other aspects of the learning process. For external-addressed performance goals, small to moderate regression coefficients were observed, indicating a differential pattern with respect to the different addressee groups. Similar to performance avoidance goals (Ames & Archer, 1988; Wolters, 2004), parent-addressed goals correlated negatively with mastery goals, academic self-concept and task value, and positively with test anxiety. With respect to the population here under investigation, eighth and ninth graders, it appears as though performance goals that are oriented on appearances vis-à-vis one's parents are less adaptive. Although the significance of parents in scholastic learning processes is common knowledge, this finding demonstrates that their relevance to goal setting processes is widely underexplored. This is also reflected in existing measurement instruments for the assessment of achievement goals, which only seldom incorporate parent-addressed goals. One implication of the present study is that measurements of goals should increasingly

focus on parents as a relevant social factor for performance goals. A more adaptive pattern resulted in conjunction with performance goals addressing classmates: These were positively associated with several facets of an adaptive learning process, namely mastery goals, effort expenditure, and task value. In contrast to performance goals addressing parents and classmates, teacher-directed goals were associated more weakly, and in a contradictory manner, according to the criteria assessed in our study. A positive association with depth of learning processes and a negative association with task value were the only significant effects, but were, with respect to direction, inconsistent with one another. One reason for this pattern might be that although teachers define learning tasks and evaluation standards, their rewards and punishments may often be of less relevance for the average 15-year-old than standards defined by peers or parental rewards and punishments (Harris, 1995).

In terms of relationships with external criteria, self-addressed goals may define a separate class of performance goals: In comparison to other addressee-specific goals, especially parent-addressed goals, a contrary pattern of regression coefficients emerged. Self-addressed performance goals were positively associated with all aspects of adaptive learning and were negatively related with our indicator of maladaptive learning, namely test anxiety. A further examination of self-directed performance goals would be an interesting task for future research. Good cause for this is provided not only by the results mentioned above but also by a remarkably high positive association with mastery goals. Here one would want to clarify whether self-addressed performance goals build up, as assumed in the present work, a separate class of performance goals or alternatively whether they combine aspects of performance goals and mastery goals, for example concerning the underlying normative and absolute standards of evaluation.

As the present study was conducted in Germany, it would be desirable to replicate the results in other countries in order to examine possible cultural differences. However, when doing so readers should be aware that the items in the questionnaire may be interpreted differently by an Anglo/American population. For example, pretests showed that different wordings used to refer to academic performance such as "I am good" and "I get a good grade" are not interpreted as being different by German students but might be understood to have disparate meanings by students in

other countries. A more serious limitation of the present study is that no disclosures could be made on the genesis of the addressee specification of performance goals. The population from which we derived our sample was made up of pupils in the eighth and ninth grades of German public schools and therefore represented a relatively narrow age range. We assumed that persons in this age group have already completed the developmental process of differentiation among performance goals concerning important others. Nevertheless, very little is known about this developmental process. Although initial findings support the assumption that, with increasing age, performance goals are increasingly differentiated with respect to different addressees (Stoeger, 2002), further research to enlighten this developmental process is clearly needed.

## References

- Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*, 261–271.
- Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology, 80*, 260–267.
- Bong, M. (2001). Between- and within-domain relations of academic motivation among middle and high school students: Self-value, task value, and achievement goals. *Journal of Educational Psychology, 93*, 23–34.
- Byrne, B. M. (1998). *Structural Equation Modeling With LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming*. Mahwah, NJ: Erlbaum.
- Church, M. A., Elliot, A. J., & Gable, S. L. (2001). Perceptions of classroom environment, achievement goals, and achievement outcomes. *Journal of Educational Psychology, 93*, 43–54.
- DeBacker Roedel, T., Schraw, G., & Plake, B. S. (1994). Validation of a measure of learning and performance goal orientation. *Educational and Psychological Measurement, 54*, 1013–1021.
- Dresel, M. (2001). A longitudinal analysis of Dweck's motivational-process-model in the classroom. *Psychology Science, 43*, 129–152.
- Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist, 41*, 1040–1048.
- Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. Philadelphia: Psychology Press.
- Dweck, C. S., & Elliott, E. S. (1983). Achievement motivation. In E. Hetherington (Ed.), *Socialisation, personality, and social development* (pp. 643–691). New York: Wiley.
- Dweck, C. S., & Henderson, V. L. (1988). *Theories of intelligence: Background and measures*. Unpublished manuscript, University of Illinois at Champaign-Urbana.
- Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist, 34*, 169–189.
- Elliot, A. J., & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology, 72*, 218–232.
- Elliot, A. J., & Harackiewicz, J. M. (1996). Approach and avoidance goals and intrinsic motivation: A mediational analysis. *Journal of Personality and Social Psychology, 70*, 461–475.
- Elliot, A. J., & McGregor, H. (2001). A 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology, 80*, 501–519.
- Elliot, A. J., & Sheldon, K. M. (1997). Avoidance achievement motivation: A personal goals analysis. *Journal of Personality and Social Psychology, 73*, 171–185.
- Gold, A., & Souvignier, E. (2000, September). *Lernstrategien und Lernerfolg* [Learning strategies and learning achievement]. Poster session presented at the 42nd Conference of the German Research Association, Jena, Germany.
- Grusec, J. E., & Goodnow, J. J. (1994). Impact of parental discipline methods on the child's internalization of values: A reconceptualization of current points of view. *Developmental Psychology, 30*, 4–19.
- Harackiewicz, J. M., Barron, K. E., Pintrich, P. R., Elliot, A. J., & Thrash, T. M. (2002). Revision of achievement goal theory: Necessary and illuminating. *Journal of Educational Psychology, 94*, 638–645.
- Harackiewicz, J. M., Barron, K. E., Tauer, J. M., Carter, S. M., & Elliot, A. J. (2000). Short-term and long-term consequences of achievement goals: Predicting interest and performance over time. *Journal of Educational Psychology, 92*, 316–330.
- Harris, J. R. (1995). Where is the child's environment? A group socialization theory of development. *Psychological Review, 102*, 458–489.
- Jöreskog, K., & Sörbom, D. (2001). *LISREL 8.51* [Computer software]. Chicago: Scientific Software International.
- Kaplan, A., & Middleton, M. J. (2002). Should childhood be a journey or a race? Response to Harackiewicz et al. (2002). *Journal of Educational Psychology, 94*, 646–648.
- Kinderman, T. A. (1993). Natural peer groups as contexts for individual development: The case of children's motivation in school. *Developmental Psychology, 29*, 970–977.
- Linnenbrink, E. A. (2005). The dilemma of performance-approach goals: The use of multiple goal contexts to promote students' motivation and learning. *Journal of Educational Psychology, 97*, 197–213.
- Maehr, M. L. (1989). Thoughts about motivation: In C. Ames & R. Ames (Eds.), *Research on motivation in education* (Vol. 3, pp. 299–315). New York: Academic Press.
- Maehr, M. L., & Midgley, C. (1991). Enhancing student motivation: A school-wide approach. *Educational Psychologist, 26*, 399–427.
- Meece, J. L., Blumenfeld, P. C., & Hoyle, R. H. (1988). Students' goal orientations and cognitive engagement in classroom activities. *Journal of Educational Psychology, 80*, 514–523.
- Middleton, M. J., & Midgley, C. (1997). Avoiding the demonstration of lack of ability: An underexplored aspect of goal theory. *Journal of Educational Psychology, 89*, 710–718.
- Midgley, C., Kaplan, A., & Middleton, M. (2001). Performance-approach goals: Good for what, for whom, under what circumstances, and at what cost? *Journal of Educational Psychology, 93*, 77–86.
- Midgley, C., Maehr, M. L., Hicks, L. H., Roeser, R. W., Urdan, T. C., Anderman, E. M., & Kaplan, A. (1996). *Patterns of adaptive learning survey*. Ann Arbor: University of Michigan.
- Miller, R. B., Greene, B. A., Montalvo, G. P., Ravindran, B., & Nichols, J. D. (1996). Engagement in academic work: The role of learning goals, future consequences, pleasing others, and perceived ability. *Contemporary Educational Psychology, 21*, 388–422.
- Pintrich, P. R. (2000). Multiple goals, multiple pathways. The role of goal orientation in learning and achievement. *Journal of Educational Psychology, 92*, 544–555.
- Pintrich, P. R., & Schunk, D. H. (2002). *Motivation in education: Theory, research and application* (2nd ed.). Englewood Cliffs, NJ: Merrill Prentice Hall.
- Schober, B., Ziegler, A., & Dresel, M. (2001). *Skalen zur Erfassung der Motivationalen Orientierung im Fach Mathematik* [Scales to assess motivational orientation in the subject of mathematics]. Unpublished manuscript, University of Munich, Germany.
- Skaalvik, E., M. (1997). Self-enhancing and self-defeating ego orientation: Relations with task and avoidance orientation, achievement, self-perceptions, and anxiety. *Journal of Educational Psychology, 89*, 71–81.
- Stoeger, H. (2002). *Soziale Performanzziele im Schulischen Leistungskontext* [Social performance goals in the school context]. Berlin: Logos.
- Urdan, T., & Mestas, M. (2006). The goals behind performance goals. *Journal of Educational Psychology, 98*, 354–365.
- Vrugt, A., Oort, F. J., & Zeeberg, C. (2002). Goal orientations, perceived



self-efficacy and study results amongst beginners and advanced students. *British Journal of Educational Psychology*, 72, 385-397.

Wentzel, K. R. (1998). Social support and adjustment in middle school: The role of parents, teachers, and peers. *Journal of Educational Psychology*, 90, 202-209.

Wentzel, K. R. (1999). Social-motivational processes and interpersonal relationships: Implications for understanding motivation at school. *Journal of Educational Psychology*, 91, 76-97.

Wolters, C. A. (1998). Self-regulated learning and college students' regulation of motivation. *Journal of Educational Psychology*, 90, 224-235.

Wolters, C. A. (2004). Advancing achievement goal theory: Using goal structures and goal orientations to predict students' motivation, cognition, and achievement. *Journal of Educational Psychology*, 96, 236-250.

Ziegler, A., Dresel, M., Schober, B., & Stoeger, H. (2005). *Ulm Motivational Test Battery (UMTB): Documentation of Items and Scales* (Ulm Educational Psychological Research Report, No. 15). Ulm, Germany: Ulm University, Department of Educational Psychology.

Ziegler, A., & Stoeger, H. (2002). Motivationale Ziele im Mathematikunterricht von MittelstufenschuelerInnen am Gymnasium [Motivational goals of secondary level students in Mathematics classes]. *Empirische Paedagogik*, 16, 57-78.

Received January 9, 2007

Revision received December 6, 2007

Accepted December 29, 2007 ■



## AMERICAN PSYCHOLOGICAL ASSOCIATION SUBSCRIPTION CLAIMS INFORMATION

Today's Date: \_\_\_\_\_

We provide this form to assist members, institutions, and nonmember individuals with any subscription problems. With the appropriate information we can begin a resolution. If you use the services of an agent, please do NOT duplicate claims through them and directly to us. **PLEASE PRINT CLEARLY AND IN INK IF POSSIBLE.**

PRINT FULL NAME OR KEY NAME OF INSTITUTION \_\_\_\_\_

MEMBER OR CUSTOMER NUMBER (MAY BE FOUND ON ANY PAST ISSUE LABEL) \_\_\_\_\_

ADDRESS \_\_\_\_\_

DATE YOUR ORDER WAS MAILED (OR PHONED) \_\_\_\_\_

CITY \_\_\_\_\_

STATE/COUNTRY \_\_\_\_\_

ZIP \_\_\_\_\_

\_\_\_\_ PREPAID \_\_\_\_ CHECK \_\_\_\_ CHARGE

CHECK/CARD CLEARED DATE: \_\_\_\_\_

YOUR NAME AND PHONE NUMBER \_\_\_\_\_

(If possible, send a copy, front and back, of your cancelled check to help us in our research of your claim.)

ISSUES: \_\_\_\_ MISSING \_\_\_\_ DAMAGED

TITLE \_\_\_\_\_

VOLUME OR YEAR \_\_\_\_\_

NUMBER OR MONTH \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

*Thank you. Once a claim is received and resolved, delivery of replacement issues routinely takes 4-6 weeks.*

(TO BE FILLED OUT BY APA STAFF)

DATE RECEIVED: \_\_\_\_\_

DATE OF ACTION: \_\_\_\_\_

ACTION TAKEN: \_\_\_\_\_

INV. NO. & DATE: \_\_\_\_\_

STAFF NAME: \_\_\_\_\_

LABEL NO. & DATE: \_\_\_\_\_

Send this form to APA Subscription Claims, 750 First Street, NE, Washington, DC 20002-4242

**PLEASE DO NOT REMOVE. A PHOTOCOPY MAY BE USED.**

# Achievement Goals and Achievement During Early Adolescence: Examining Time-Varying Predictor and Outcome Variables in Growth-Curve Analysis

S. Serena Shim  
Northern Arizona University

Allison M. Ryan and Carolyn J. Anderson  
University of Illinois at Urbana–Champaign

The present study advances understanding of (a) the development of achievement goals, (b) the changing association of achievement goals and achievement over time, and (c) the implications of changes in achievement goals for changes in achievement over time. African American and European American adolescents' ( $N = 588$ ) achievement goals and subsequent achievement were assessed at 4 time points (fall and spring of 6th and 7th grades) and modeled using growth-curve analytic techniques. There was an overall decline in all 3 types of achievement goals (mastery, performance-approach, and performance-avoidance goals), because of within-year rather than between-year decreases. The association between mastery goals and achievement was null at Time 1 and then positive at the following 3 time points. The association between performance-approach goals and achievement went from negative to null across time. Changes in students' goals, as well as their initial levels of goals, were particularly important in understanding how mastery goals foreshadow achievement. The implications of the findings for both theory and practice are discussed.

**Keywords:** achievement goal orientation, middle-school transition, adolescence, motivation, achievement

A prominent construct in current theoretical models of motivation is the achievement goal (Wigfield, Eccles, Schiefele, Roeser, & Davis-Kean, 2006). Research using an achievement-goal framework has flourished in recent years. Extant work has primarily focused on the relation of achievement goals to a variety of educationally relevant outcomes and the influence of different instructional contexts on achievement goals (Wigfield et al., 2006). There has been little longitudinal research on achievement goals. Thus, our understanding of how achievement goals change across time and the implications for students' achievement across time is incomplete. Therefore, in this article, we examine the developmental processes of achievement goals and achievement with a longitudinal research design and growth-curve analytic techniques that incorporate changes in the predictor variables (i.e., achievement goals) as well as the outcome variable (i.e., achievement). We begin with an overview of theory and research on the achievement-

goal construct, highlighting why a longitudinal analytical approach is appropriate and will advance understanding of the nature of achievement goals and the consequences of achievement goals for achievement. We then discuss our approach and rationale for the three key questions that guide our longitudinal investigation: (a) What is the nature of the development of achievement goals? (b) Does the association between achievement goals and achievement change across time? and (c) Do changes in achievement goals have implications for changes in achievement?

## Overview of Theory and Research on Achievement Goals

Achievement goals represent different orientations toward academic competence that students often have in achievement settings (see Elliot, 2005, for a review). Researchers have distinguished between a mastery goal (a focus on developing academic competence) and performance goals (a focus on demonstrating academic competence to others, especially via social comparisons of relative ability; e.g., Ames & Archer, 1988; Dweck & Leggett, 1988; Harackiewicz & Elliot, 1993; Nicholls, 1989). Researchers have found it is important to distinguish performance goals as approach oriented or avoidant oriented (Elliot & Church, 1997; Middleton & Midgley, 1997; Skaalvik, 1997). A performance-approach goal concerns a focus on demonstrating high competence and gaining positive judgments from others. A performance-avoidance goal concerns a focus on avoiding the demonstration of incompetence and preventing negative judgments from others. For example, a student with a mastery goal might reflect on such questions as "Am I learning? Have I improved my skills?"; a student with a performance-approach goal might reflect on such questions as "Do I look smart? Did I do better than others?"; and a student with a performance-avoidance goal might reflect on such questions as "Do I look dumb? Did I perform worse than others?"

---

S. Serena Shim, Department of Educational Psychology, College of Education, Northern Arizona University; Allison M. Ryan and Carolyn J. Anderson, Department of Educational Psychology, College of Education, University of Illinois at Urbana–Champaign.

This article was based on a doctoral dissertation submitted by S. Serena Shim to the University of Illinois at Urbana–Champaign. We thank Eva Pomerantz and Phil Rodkin for the helpful suggestions and insightful comments they offered as part of S. Serena Shim's dissertation committee.

Correspondence concerning this article should be addressed to S. Serena Shim, Department of Educational Psychology, College of Education, Northern Arizona University, P.O. Box 5774, Flagstaff, AZ 86011, or to Allison M. Ryan, Department of Educational Psychology, College of Education, University of Illinois at Urbana–Champaign, 1310 South Sixth Street, Champaign, IL 61820. E-mail: serena.shim@nau.edu or ryan2@uiuc.edu



Different goals set in motion disparate cognitive, emotional, and behavioral processes (Dweck & Leggett, 1988; Elliot, 2005). Achievement goals are viewed as a proximal influence on competence-relevant processes and outcomes; they are a precursor to behavior (i.e., why a student does something precedes whether, if, and how they actually do it; Elliot & Church, 1997). Personality or dispositional factors are postulated to partially explain individual differences in achievement goals (e.g., Dweck & Leggett, 1988; Elliot, 1999; Harackiewicz, Barron, & Elliot, 1998). However, theorists describe achievement goals as fluctuating in relation to different experiences in learning environments (Ames, 1992; Maehr & Midgley, 1996). Experimental work has been successful in orienting students to pursue different achievement goals (e.g., Butler, 1993; Elliot & Harackiewicz, 1996; Newman & Schwager, 1995). Observational work that links classroom experiences to personal goals is also evidence for this viewpoint (Linnenbrink, 2005; Patrick, Anderman, & Ryan, 2002; Turner, Meyer, & Midgley, 2003). Thus, achievement goals are conceptualized as dynamic and changing in relation to features of the context.

The analytic methods used to study the effects of goals have not captured the changing nature of achievement goals. The vast majority of quantitative research on the effects of achievement goals (i.e., goals as predictors) falls into one of three designs: (a) goals and outcomes measured together at one point in time (e.g., Middleton & Midgley, 1997; A. M. Ryan & Pintrich, 1997; Wolters, 2004), (b) goals measured at Time 1 and outcomes measured at Time 2 (e.g., Elliot, McGregor, & Gable, 1999; Grant & Dweck, 2003; Harackiewicz, Barron, Tauer, & Elliot, 2002), and (c) goals measured at Time 1 and outcomes measured at Time 1 and Time 2 (e.g., McGregor & Elliot, 2002; Senko & Harackiewicz, 2005; Shim & Ryan, 2005; Stipek & Gralinski, 1996). We propose that growth-curve analytic techniques that incorporate the changing nature of achievement goals are well suited to theoretical conceptualizations of achievement goals and provide a more complete understanding of the development of achievement goals and the consequences for achievement.

### What Is the Nature of the Development of Achievement Goals?

Our first concern is to understand the nature of the development of achievement goals during early adolescence. Prior longitudinal research examining the development of achievement goals during early adolescence (i.e., goals as outcomes) has assessed achievement goals at two time points and found that a mastery goal declines and a performance-approach goal increases across the transition to middle school (E. M. Anderman & Midgley, 1997; L. H. Anderman & Anderman, 1999).<sup>1</sup> However, analyzing change across two time points provides a limited understanding of the development of achievement goals. At least three time points are needed to document a trend (Rogosa, 1988). Further, these studies have assessed achievement goals at one point in different grades, and thus, within- and between-school-year changes have not been disentangled. Changes in motivation over time within the same classroom environment may be different from changes that occur as students move into new classrooms or school environments. Such information is particularly important regarding the transition to middle school, given the long-standing concerns that middle-level schools are a major source of declines in motivation during

early adolescence (see Eccles, Lord, Roeser, Barber, & Jozefowicz, 1997; Juvonen, Le, Kaganoff, Augustine, & Constant, 2004). Are there normative declines in achievement goals in progress prior to the transition? Are declines in achievement goals seen in the beginning of the middle-school year, or do they unfold during the first year in middle school? The present study addresses such unanswered questions and provides a more complete understanding of the nature of change in goals during early adolescence by estimating growth trajectories across four time points (data were collected in 6-month increments in the fall and spring of both sixth and seventh grades, spanning the transition to middle school). We examine the overall trend in development across the four time points as well as investigate the within- and between-year patterns of change that contribute to any documented trends.

We subsequently examine interindividual differences to see whether growth trajectories vary across individuals and, if so, whether some of this variation can be explained by gender and/or race. Findings regarding gender differences in achievement goals are not entirely consistent, but when differences are found, girls tend to be more mastery oriented than boys (e.g., Ablard & Lipschultz, 1998; Freeman & Anderman, 2005; Kenney-Benson, Pomerantz, Ryan, & Patrick, 2005; Meece & Holt, 1993; Nolen, 1988), and boys tend to be more performance oriented than girls (e.g., L. H. Anderman & Anderman, 1999; Middleton & Midgley, 2002; Roeser, Midgley, & Urdan, 1996; A. M. Ryan, Hicks, & Midgley, 1997; Stipek & Gralinski, 1996). There has been far less research on race differences in achievement goals, but some research suggests that African American students may be higher in mastery (e.g., Kaplan & Maehr, 1999; Middleton & Midgley, 2002) as well as performance-approach and performance-avoidance goals (Middleton & Midgley, 2002). However, given the limited longitudinal analyses of achievement goals, we do not know if gender and race relate to the development of goals across time or if the relation of gender and race to achievement goals changes across time. The relations of gender and race might differ when students are in elementary versus middle school, as both gender and ethnicity take on new significance. Gender and race differences may emerge or grow stronger as students cognitively mature and reflect on their identity in more complex ways during adolescence.

### Does the Association Between Achievement Goals and Achievement Change Over Time?

Prior research on achievement goals and achievement has yielded inconsistent findings and spurred debate as to the consequences of achievement goals for achievement (Harackiewicz, Barron, Pintrich, Elliot, & Thrash, 2002; Midgley, Kaplan, & Middleton, 2001). The consequences of achievement goals may change as students mature and move through the educational system (see Midgley et al., 2001). Growth-curve models that treat

<sup>1</sup> Several other studies have examined the relation of achievement goals across two time points (e.g., Middleton, Kaplan, & Midgley, 2004; or Gehlbach, 2006), but the focus was on stability across time (i.e., correlations, beta coefficients), not normative developmental trends (i.e., mean level differences between the two time points were not reported). Because our study focuses on developmental patterns, we do not review these stability coefficients here.



achievement goals as time-varying predictors are well suited to examine the changes in the achievement goal–achievement relation. To examine the implications of achievement goals for subsequent student achievement, we incorporate the measurements of achievement goals at all four time points into our models for achievement and test whether the effects of goals on subsequent achievement are consistent over time. Because a performance-avoidance goal has consistently been found to be negatively associated with achievement, our focus is on the associations between mastery and performance-approach goals and achievement. When effects are found, a mastery goal is positively related to achievement for younger students (e.g., Kaplan & Maehr, 1999; Midgley & Urdan, 1995; Roeser et al., 1996; A. M. Ryan, Patrick, & Shim, 2005; Wentzel, 1993; Wolters, Yu, & Pintrich, 1996; but see also Pajares, Britner, & Valiante, 2000; Pajares & Valiante, 2001; Pintrich, 2000; Skaalvik, 1997, for studies documenting no relation).<sup>2</sup> A mastery goal has consistently been found to have no association with achievement for college students (see Harackiewicz, Barron, Pintrich, et al., 2002, for a review). A performance-approach goal has been found to have no association with achievement for younger students (e.g., Pajares et al., 2000; Pajares & Valiante, 2001; but see A. M. Ryan et al., 2005, for a study documenting a negative relation). However, a performance-approach goal has consistently been found to have a positive association with achievement for college students (see Harackiewicz, Barron, Pintrich, et al., 2002, for a review). As students advance through the educational system, the context becomes more competitive (e.g., grading on a curve), and a mastery goal may be less adaptive, whereas a performance-approach goal may be more adaptive for achievement. The transition to middle school represents such a change in context (see Eccles, 2005, for a review), and thus, our longitudinal study spanning this transition is at an appropriate time to examine whether the association of achievement goals to achievement changes.

### Do Changes in Achievement Goals Have Implications for Changes in Achievement?

Motivational research most typically emphasizes interindividual effects, compared with intraindividual effects. Interindividual effects concern variations between students and document if high or low scores on an aspect of motivation, compared with other students, explain variation in engagement or achievement. Intraindividual effects concern the changes in individuals' motivation that occur across time and document if such within-individual fluctuations are important for changes in engagement or achievement. Our first two research questions concern interindividual analyses and identify general patterns across students (the first pertaining to the development of goals and the second pertaining to the associations of goals to achievement across time). However, we expect that there is variability in those general patterns between students. For example, we may document a general decline in mastery goals over time, but not all students will conform to this general pattern. Our third research question concerns intraindividual analyses and examines the implications for achievement of varying within-individual patterns of change in achievement goals. Examining intraindividual effects contributes to a more complete understanding of how motivational beliefs may support or hinder achievement across time.

Prior research has incorporated intraindividual change in achievement goals using residual scores (Time 2 goal scores are regressed on Time 1 goal scores; see Meece & Miller, 2001). Another approach used was the creation of change scores (goal scores at Time 2 minus goal scores at Time 1; see Gehlbach, 2006). Such studies concluded that when a mastery goal declines, various aspects of engagement, as well as achievement, decline, and when a mastery goal increases, engagement and achievement increase. Another similar study examined change scores in students' perceptions of teacher encouragement of different achievement goals and found that when students' perceptions of their teacher emphasizing mastery goals declined from one year to the next, their engagement and achievement declined (Urden & Midgley, 2003). No effects for performance goals were found in any of these studies; however, performance-approach and performance-avoidance goals were not distinguished (Gehlbach, 2006; Urden & Midgley, 2003) or not examined (Meece & Miller, 2001).

Such research recognizes the importance of intraindividual change in achievement goals but does not provide a complete understanding of the implications of changes in students' achievement goals. These analytic approaches examine the direction of change but do not incorporate the general level of achievement goals during the two time periods (e.g., in Urden & Midgley's 2003 study, approximately three fifths of students were categorized as "no-change," and within this group were students whose goal scores were high, medium, or low at both time points). In contrast, we use growth-curve models to capture both the level of achievement goals and the pattern of change within students across four time points. We include two predictors: students' achievement-goal scores at Time 1 (termed baseline score) and deviations from their Time 1 achievement-goal scores (termed *DevT<sub>i</sub>* score). A baseline score, the most typical predictor examined in longitudinal analyses indicates whether students' achievement goals at the beginning of the study predict the growth trajectory in achievement over time. The *DevT<sub>i</sub>* scores indicate whether gains and losses from Time 1 scores are also important. By including both predictors, we are able to ascertain whether changes in achievement goals, beyond students' initial levels of achievement goals, contribute to an understanding of students' achievement trajectories during early adolescence.

### Overview of the Present Research

In summary, in the present study, we use a longitudinal design and growth-curve analytic techniques to advance understanding of

<sup>2</sup> Achievement refers to student grades, not standardized tests. This is appropriate to compare patterns between younger students in public schools and college students, as college students do not routinely take standardized tests, and thus, research on college students' achievement goals and achievement has examined grades as a measure of achievement. In addition, the present study uses grades as an indicator of student achievement, so it makes sense to focus the literature review in this way. Further, in our review of the relationship between achievement goals and achievement, we examine the zero-order correlations, not the unique effects of goals above and beyond other achievement-related variables in multiple regression models. These types of multiple regression models vary greatly across studies and include variables that may be mediators of the relationship between achievement goals and achievement, which is not the focus of the present study.



(a) the normative patterns of development of achievement goals, (b) the potentially changing association of achievement goals and achievement over time, and (c) the implications of within-individual changes in achievement goals (with attention to both direction of change and level of goals) for changes in achievement over time. Our longitudinal design includes four measurements of achievement goals in the fall and spring of sixth and seventh grades. Consistent with our conceptualization that achievement goals predict achievement, our four measurements of achievement (i.e., grade-point average [GPA] computed from report card grades) were collected at the end of the corresponding semesters and thus encompass achievement that extends beyond the measure of achievement goals. When we examined the associations of achievement goals to achievement across time, we controlled for prior achievement (i.e., fifth-grade GPA) because relations between achievement goals and achievement are reciprocal and we wanted to examine the effects of achievement goals on achievement above and beyond the contribution of prior achievement history (see Brophy, 2005; Senko & Harackiewicz, 2005, for a discussion of these issues).

## Method

### Procedure

The data were collected as part of the University of Illinois at Urbana-Champaign Adolescent Transitions Project, which is a 2-year longitudinal study examining changes in academic and social adjustment across the transition to middle school. Participants attended 1 of 15 elementary schools when they were in sixth grade and moved into one of three middle schools when they were in seventh grade. These predominantly low-income schools (the average rate of eligibility for free or reduced fee lunch was 66% across the elementary schools and 59% across the middle schools) served nonmetropolitan, small urban communities.

Letters describing the project were given to all students to take home to their parents 2 weeks prior to each data collection. The average mobility rate (average percentage of students who transfer in or out of the school within an academic year) in these schools was high (22% across the elementary schools and 24% across the middle schools), and thus, we maximized the sample size by recruiting new students at each time point. If parents did not want their children to participate in the study, they were instructed to have their child return an attached form to the teacher, to call the school, or to call the researchers at the university number provided on the letter. Fewer than 5% of the parents declined to have their child participate at any time point.

Surveys were administered to students in their classrooms. Instructions and items were read aloud while students read along and responded. Students were told that the purpose of the survey was to find out about students' beliefs and behaviors and that completing it was voluntary. Students were assured that the information in the survey would be kept confidential. We visited the schools on 1 additional day to administer make-ups for students who were absent for survey administration.

### Participants

A total of 738 students participated in the study at Time 1 and had complete information regarding their achievement goals and

achievement. At subsequent time points, we lost some students from our sample (88, 182, and 110, at Times 2, 3, and 4, respectively) and gained some new students (110, 280, and 40, at Times 2, 3, and 4, respectively). Sample instability was due to two factors: school participation and mobility rates. Regarding school participation, we lost some students who went to a nonparticipating middle school and gained some students from nonparticipating elementary schools (in total, three elementary schools and one middle school chose not to participate). After these students were accounted for, the sample instability was comparable with the mobility rates reported by the state for these schools. Because we were examining developmental changes over the middle-school transition, only those students who actually made the transition into middle school (i.e., had data for at least one time point in elementary school and at least one time point in middle school) were included in the analyses. Students who were not African American or European American were dropped because there were too few ( $n = 30$ ) to examine ethnic differences. These restrictions yielded a sample of 588 students (56% girls, 44% boys; 60% African American, 40% European American). There were no significant differences between students in our longitudinal sample and students excluded from our longitudinal sample (i.e., had data at only one of the time points) on achievement goals or achievement at any of the four time points.

### Measures

**Achievement goals.** We used the Patterns of Adaptive Learning Survey (Midgley et al., 1997) to assess achievement goals. *Mastery-approach goal* items (six items) refer to a focus on developing academic competence (e.g., "An important reason I do my work is because I want to improve my skills"). *Mastery-avoidance goals* are not measured in the present study, and thus, mastery-approach goals are referred to as *mastery goals* hereafter. *Performance-approach goal* items (five items) refer to a focus on demonstrating high academic competence relative to other students in the class (e.g., "I like to show my teacher that I'm smarter than the other students in my class"). *Performance-avoidance goal* items (five items) concern a focus on avoiding looking inferior relative to other students in the class ("An important reason I do my work is so that the teacher doesn't think I know less than others"). Goals were found to be reliable in the present sample (see Table 1 for alpha coefficients).

**Achievement.** Students' grades in reading, math, science, English, and social studies were collected from their school records. The grades were coded 1 (*F*) through 13 (*A+*). We computed a GPA for each semester by taking the mean of the five subject grades.

### Handling Missing Data

Hierarchical linear modeling (HLM) is flexible in handling missing data at Level 1, but it cannot handle missing values at Level 2. In the current study, we control for prior GPA and use Time 1 achievement goals to calculate the deviation from Time 1 scores ( $DevT_1$ ). Thus, we had four important Level 2 variables (fifth-grade GPA and three Time 1 achievement goals), and students who joined the study after Time 1 did not have these variables. To include the maximum number of students in the

Table 1  
Means, Standard Errors, and Reliabilities of Achievement Goals and GPA

Time	Mastery goals			Performance-approach goals			Performance-avoidance goals			GPA	
	<i>M</i>	<i>SE</i>	$\alpha$	<i>M</i>	<i>SE</i>	$\alpha$	<i>M</i>	<i>SE</i>	$\alpha$	<i>M</i>	<i>SE</i>
1	3.59	.04	.79	3.34	.04	.75	3.29	.05	.72	7.82	.11
2	3.42	.04	.80	3.00	.05	.78	2.92	.05	.75	7.93	.11
3	3.41	.04	.78	3.04	.05	.80	2.89	.05	.66	6.74	.13
4	3.30	.04	.82	2.95	.05	.82	2.77	.05	.67	6.19	.13

Note. GPA = grade point average. Means are averaged from five data sets through PROC MIANALYZE (SAS, 2003).

analyses, missing values for these Level 2 variables were imputed. A fundamental weakness of single value imputation techniques (e.g., mean substitution, person mean substitution, regression imputation) is the underestimation of standard errors, which affects Type I errors (Allison, 2002; Newman, 2003). To avoid this problem, we employed a multiple imputation (MI) method to fill in the missing data. A strength of MI estimate is the decreased sensitivity to violations of multivariate normality (Allison, 2002). In addition, the MI paradigm neither requires nor assumes that nonresponse is ignorable (Schafer, 1997). MI has been used in HLM contexts (Dearing, Kreider, Simpkins, & Weiss, 2006) to deal with missing predictors at Level 2. In MI, a complete data set is created by filling in a random value for each missing data point. This procedure is repeated to create multiple complete data sets, each of which is subsequently analyzed.

Given that 16% of the participants had one or more missing values, we created five independent data sets to obtain 95% of efficiency (Rubin, 1987, p. 114).<sup>3,4</sup> Then correlations, *t* tests, and HLM analyses were performed on the five imputed datasets. We used PROC MIANALYZE (SAS, 2003) to combine the results from these five data sets and generate valid statistical inference of the parameters. PROC MIANALYZE provides a *t* test, taking into consideration both within- and between-imputation variance.

## Results

### Analyses and Model Specification

The estimated means, standard errors, and reliability coefficients for achievement goals and GPA are provided in Table 1. The presented means are estimated based on the five imputed datasets. Our growth trajectories were fitted to data using PROC MIXED in SAS (Version 9.1; SAS, 2003) with maximum-likelihood estimation.<sup>5</sup> Robust standard errors (which are less biased, even when the random structure of the model is misspecified) were used for testing the fixed effects. All models were simplified until all fixed effects reached significance ( $p < .05$ ). Gender, race, and their interaction terms were tested in preliminary analyses and retained in the final models only when significant. When it yielded a better model fit to data, gender and racial group membership with four categories was used instead of gender or race separately.

### What Is the Nature of the Development of Achievement Goals?

To estimate the general trend of the development of achievement goals across four time points, we estimated three growth-

curve models with each achievement goal as an outcome variable. We estimated the following growth-curve model to examine the general changes in each achievement goal across the four time points:

$$\text{Level 1: Goal}_{ij} = \beta_{0j} + \beta_{1j}(\text{Time})_{ij} + \varepsilon_{ij}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + U_{0j}$$

$$\beta_{1j} = \gamma_{10} + U_{1j},$$

where  $\beta_{0j}$  and  $\beta_{1j}$  are the random intercepts and slopes, respectively, and  $\gamma_{00}$  and  $\gamma_{10}$  are the fixed intercepts and slopes, respectively.

The estimated intercepts, which are the average starting points for the population, were 3.57 for mastery goals (95% confidence interval [CI] = 3.49, 3.64), 3.25 for performance-approach goals (95% CI = 3.17, 3.34), and 3.21 for performance-avoidance goals (95% CI = 3.11, 3.30), respectively. All three types of achievement goals showed a general decline over time, as indicated by the estimated fixed effects of time in each of the three models. Mastery goals declined by  $-.09$  (95% CI =  $-.12$ ,  $-.06$ ). Performance-approach goals declined by  $-.11$  (95% CI =  $-.15$ ,  $-.08$ ). Performance-avoidance goals declined by  $-.16$  (95% CI =  $-.20$ ,  $-.12$ ). There was significant individual variability in both the initial levels and rates of changes of the growth trajectories of all three types of achievement goals, as indicated by the random parts of the three models (see Table 2).

There were substantively meaningful distinctions between each of four time points (between Times 1 and 2, children had one teacher in an elementary-school classroom; between Times 2 and 3, children made the transition to a new middle-school environment; and between Times 3 and 4, children rotated amongst different teachers in middle school). We conducted a series of *t*

<sup>3</sup> We imputed missing values in all variables to do descriptive analyses, *t* tests, and correlations. Level 1 variables do not need to be imputed to conduct HLM.

<sup>4</sup> We ran the analyses with the complete data subset (i.e., the subset of students for whom there were no missing data) without imputing missing values, and the conclusions were identical.

<sup>5</sup> Unstructured variance-covariance matrices for the intercepts and slopes were fit. Various Level 1 error covariance matrices, such as autocorrelated errors, moving average, and autocorrelated moving average were examined in the preliminary analyses, but the inclusion of additional parameters was not warranted by likelihood-ratio test. Thus,  $\sigma^2 I$  (constant and uncorrelated Level 1 error structures) was used for error covariance matrices.



tests to examine changes in goals across different time points (see Table 3 for the results). The results have shown significant mean differences in all three achievement goals between Times 1 and 2 and between Times 3 and 4. There were no significant mean differences in any of the three achievement goals between Times 2 and 3. Thus, declines were more pronounced within, rather than between, school years. Declines within the school year occurred in both elementary and middle school. The declines that occurred between the end of elementary and the beginning of middle school were not significant, indicating that the middle-school transition was not the major contributing source to the overall pattern of decline found for achievement goals during early adolescence.

Next, we examined whether gender and/or race explained variation in students' growth trajectories in achievement goals (see Table 4 for the results). We entered gender and race as Level 2 predictors in the growth-curve models to examine whether the initial levels (i.e., the intercept) and rates of change (i.e., the slope) of achievement goals are a function of these variables. The HLM equation to estimate the gender and racial differences in the changes in achievement goals across the four measurements takes the following form:

$$\text{Level 1: Goal}_{ij} = \beta_{0j} + \beta_{1j}(\text{Time})_{ij} + \epsilon_{ij}$$

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Gender})_j + \gamma_{02}(\text{Race})_j + \gamma_{03}(\text{Gender} \times \text{Race})_j + U_{0j}$$

Table 2  
General Trend of Changes in Achievement Goals

Model parameter	Estimate	SE	df	t
Mastery goals				
Fixed-effect parameter				
M initial level	3.57	.04	110	87.60***
M growth rate	-0.09	.02	92	-5.77***
Variance parameter				
Initial status	0.55	.05	8596	11.55***
Covariance	-0.08	.02	101	-4.60***
Linear growth rate	0.04	.01	19	4.43***
Level 1 error	0.34	.02	18	17.66***
Performance-approach goals				
Fixed-effect parameter				
M initial level	3.25	.04	479	78.45***
M growth rate	-0.11	.02	333	-7.23***
Variance parameter				
Initial status	0.58	.06	1000	10.13***
Covariance	-0.01	.02	76	-0.45
Linear growth rate	0.03	.01	19	2.88***
Level 1 error	0.48	.02	56	20.74***
Performance-avoidance goals				
Fixed-effect parameter				
M initial level	3.21	.05	267	67.56***
M growth rate	-0.16	.02	230	-8.37***
Variance parameter				
Initial status	0.65	.08	157	8.39***
Covariance	-0.05	.03	38	-1.80
Linear growth rate	0.04	.02	22	2.23*
Level 1 error	0.73	.04	25	18.73***

Note. *dfs* are from PROC MIANALYZE (SAS, 2003), not from hierarchical linear modeling.

\* $p < .05$ . \*\*\* $p < .001$ .

Table 3  
Mean Differences in Achievement Goals at Four Time Points

Comparison	M difference	95% Confidence interval	SE	df	t
Mastery goals					
Time 1-2	.17	.09, .24	.04	1867	4.39***
Time 2-3	.02	-.06, .10	.04	121	0.49
Time 3-4	.10	.03, .17	.04	112	2.90**
Performance-approach goals					
Time 1-2	.33	.24, .43	.05	553	7.25***
Time 2-3	-.04	-.12, .04	.04	99	-0.85
Time 3-4	.09	.01, .18	.04	127	2.24*
Performance-avoidance goals					
Time 1-2	.38	.25, .50	.06	68	6.15***
Time 2-3	.02	-.09, .15	.06	52	0.45
Time 3-4	.12	.02, .21	.02	233	2.35*

Note. *t*-test results from five imputed data sets were averaged through PROC MIANALYZE procedure (SAS, 2003). Thus, *dfs* are from PROC MIANALYZE, not from *t* tests.

\* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(\text{Gender})_j + \gamma_{12}(\text{Race})_j + \gamma_{13}(\text{Gender} \times \text{Race})_j + U_{1j},$$

where  $\text{Gender}_j = 0$  for boys and 1 for girls and  $\text{Race}_j = 0$  for European American and 1 for African American.

### A Mastery Goal

Gender and race were associated with the initial level but not the rate of change in mastery goals. Female students started out with higher levels of mastery goals than did male students. African American students had higher levels of mastery goals than did European American students. The finding that neither gender nor race predicted the slope means that these group differences remained constant across the four time points. The separate main effects of gender and race on the intercept indicated that African American girls were the highest, followed by European American girls, African American boys, and European American boys (all pairwise comparisons were significant). Their level of mastery goals varied, but their growth trajectories moved in parallel fashion across time (see Figure 1).<sup>6</sup>

### A Performance-Approach Goal

Gender and race were associated with both the initial level and the rate of change in performance-approach goals. There was a Gender  $\times$  Race interactive effect on the intercept and a main effect of gender on the rate of change. The results indicated that African American boys had the highest level of performance-approach goals at all time points, compared with all other students. European American boys were similar to all girls initially but had higher

<sup>6</sup> Significant random effects were found in all three goal models. The results indicated that there was significant individual variability in both intercept and slope. However, this is not the focus of the current investigation and, hence, is not discussed.

Table 4  
Gender and Race Differences in Changes in Achievement Goals

Model parameter	Estimate (95% confidence interval)	SE	df	t
Mastery goals				
Initial status				
M initial status	3.81 (3.71, 3.90)	0.05	651	77.80***
Gender	-0.17 (-.28, -.06)	0.06	46975	-3.06**
Race	-0.36 (-.47, -.24)	0.06	1227	-6.15***
Linear change				
M change rate	-0.09 (-.12, -.06)	0.02	92	-5.77***
Performance-approach goals				
Initial status				
M initial status	3.17 (3.03, 3.32)	0.07	1308	42.57***
Gender	0.31 (.10, .52)	0.11	56267	2.94**
Race	0.03 (-.17, .22)	0.10	1084	0.27
Gender $\times$ Race	-0.32 (-.60, -.03)	0.15	254	-2.17*
Linear change				
M change rate	-0.15 (-.20, -.11)	0.02	1272	-7.09***
Gender	0.09 (.02, .15)	0.03	1272	2.69**
Performance-avoidance goals				
Initial status				
M initial status	3.03 (2.91, 3.14)	0.06	506	50.96***
Gender	0.39 (.25, .53)	0.07	801	5.39***
Linear change				
M change rate	-0.16 (-.20, -.12)	0.02	230	-8.37***

Note. Omitted categories are girls and African Americans. Estimates for boys and European Americans are shown in the table. *dfs* are from PROC MIANALYZE (SAS, 2003), not from hierarchical linear modeling.

\* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

levels of performance-approach goals than did all girls later in time (see Figure 2).

#### A Performance-Avoidance Goal

Gender, but not race, predicted the initial levels of performance-avoidance goals. Boys had higher levels of performance-avoidance goals than did girls at Time 1. Neither gender nor race predicted the rate of change in performance-avoidance goals. Thus, the trajectories of performance-avoidance goals for boys and girls are parallel, but the levels of performance-avoidance goals were consistently higher among boys than girls.

#### Does the Association Between Achievement Goals and Achievement Change Across Time?

To address whether achievement goals are predictors of achievement across time, we estimated an unconditional model for GPA to ensure that there was significant random variance in the intercept and slope to proceed with modeling achievement goals as predictors of GPA. The intercept for GPA was 8.08, and the slope was  $-.61$ , indicating that GPA declined over time.

There was significant variability in both the initial level and rate of change over time, as indicated by the random parts of the model:

$$\begin{aligned}\widehat{\text{var}}(U_{0j}) &= 6.33, 95\% \text{ CI } (5.48, 7.18), \text{ and } \widehat{\text{var}}(U_{1j}) \\ &= .53, 95\% \text{ CI } (.42, .64).\end{aligned}$$

The HLM equation to estimate the effects of achievement goals on growth trajectories of achievement takes the following form:

$$\begin{aligned}\text{Level 1: } \text{GPA}_{ij} &= \beta_{0j} + \beta_{1j}(\text{Time})_{ij} + \beta_{2j}(\text{MG})_{ij} + \beta_{3j}(\text{PAP})_{ij} \\ &+ \beta_{4j}(\text{PAV})_{ij} + \beta_{5j}(\text{Time})_{ij}(\text{MG})_{ij} + \beta_{6j}(\text{Time})_{ij}(\text{PAP})_{ij} \\ &+ \beta_{7j}(\text{Time})_{ij}(\text{PAV})_{ij} + \varepsilon_{ij},\end{aligned}$$

where  $(\text{MG})_{ij}$ ,  $(\text{PAP})_{ij}$ ,  $(\text{PAV})_{ij}$  represent goal scores measured at different time points for mastery, performance-approach, and performance-avoidance goals, respectively.

$$\begin{aligned}\text{Level 2: } \beta_{0j} &= \gamma_{00} + \gamma_{01}(\text{Gender})_j + \gamma_{02}(\text{Race})_j + \gamma_{03}(\text{Gender} \\ &\times \text{Race})_j + \gamma_{04}(\text{5th Grade GPA})_j + U_{0j}\end{aligned}$$

$$\begin{aligned}\beta_{1j} &= \gamma_{10} + \gamma_{11}(\text{Gender})_j + \gamma_{12}(\text{Race})_j + \gamma_{13}(\text{Gender} \\ &\times \text{Race})_j + \gamma_{12}(\text{5th Grade GPA})_j + U_{1j}\end{aligned}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}(\text{Gender})_j + \gamma_{22}(\text{Race})_j + \gamma_{23}(\text{Gender} \times \text{Race})_j$$

For the remaining  $k = 2$  through 7,

$$\beta_{kj} = \gamma_{k0} + \gamma_{k1}(\text{Gender})_j + \gamma_{k2}(\text{Race})_j + \gamma_{k3}(\text{Gender} \times \text{Race})_j.$$

In contrast to predictors like gender or race, achievement goals do change over time. Thus, in addressing our question about whether the association between achievement goals and achievement change across time, we incorporated achievement goals at Level 1. A changing association can be captured by a significant



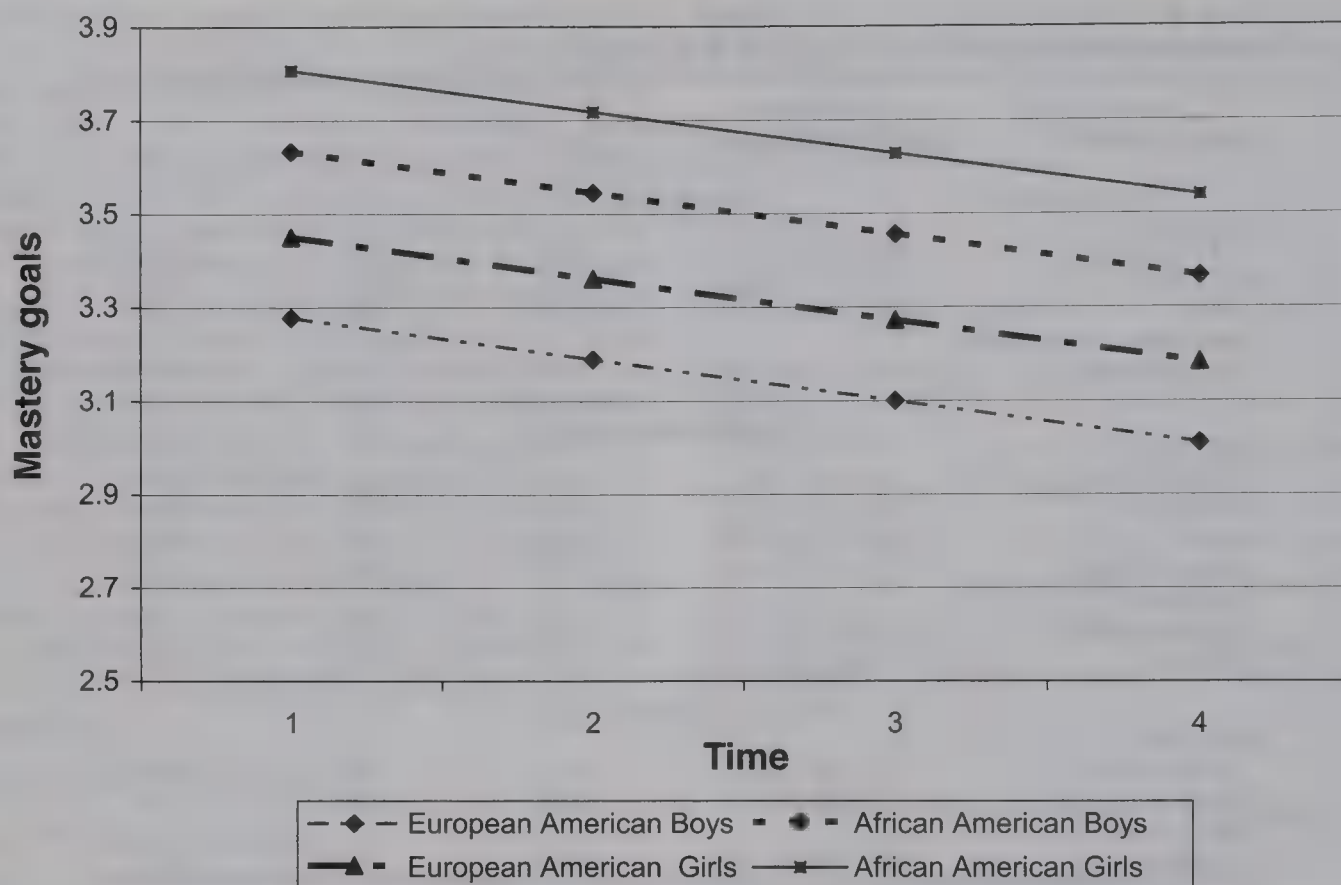


Figure 1. Gender and race differences in the development of mastery goals.

interaction between achievement goals and time.<sup>7</sup> We found significant interactions between mastery goals and time and performance-approach goals and time (see Table 5). As recommended by Singer and Willett (2003), we repeated the analyses three more times with time centered at each of the three subsequent time points so we could compare the associations of goals and achievement at different time points. For example, when time is recentered around Time 2,  $\gamma_{20}$  indicates average association between a mastery goal and GPA at Time 2, whereas when time is recentered around Time 3,  $\gamma_{20}$  indicates average association between a mastery goal and GPA at Time 3.

A mastery goal was a significant positive predictor of GPA at all time points except Time 1. The estimates for the mastery goal effect on the levels of GPA were as follows: Time 1,  $\gamma_{20} = .00$  (95% CI =  $-.13, .13$ ),  $t(193) = -.06$ ,  $p = ns$ ; Time 2,  $\gamma_{20} = .13$  (95% CI =  $.02, .24$ ),  $t(45) = 2.33$ ,  $p < .05$ ; Time 3,  $\gamma_{20} = .26$  (95% CI =  $.13, .39$ ),  $t(32) = 4.11$ ,  $p < .001$ ; and Time 4,  $\gamma_{20} = .40$  (95% CI =  $.22, .57$ ),  $t(46) = 4.57$ ,  $p < .0001$ . A performance-approach goal had an initial negative impact that dissipated over time and disappeared once students moved into middle school. The estimates for performance-approach goal effects on the levels of GPA were as follows: Time 1,  $\gamma_{30} = -.19$  (95% CI =  $-.33, -.05$ ),  $t(31) = -2.77$ ,  $p < .01$ ; Time 2,  $\gamma_{30} = -.11$  (95% CI =  $-.23, .01$ ),  $t(17) = -1.94$ ,  $p < .10$ ; Time 3,  $\gamma_{30} = -.02$  (95% CI =  $-.14, .09$ ),  $t(22) = -.43$ ,  $p = ns$ ; and Time 4,  $\gamma_{30} = .06$  (95% CI =  $-.09, .20$ ),  $t(56) = 0.81$ ,  $p = ns$ . A performance-avoidance goal was a significant negative predictor of GPA at all time points. The effect of performance-avoidance goal on the initial level of GPA was  $-.09$  (95% CI =  $-.02, -.16$ ),  $t(223) = 2.43$ ,  $p < .05$ . The direction and magnitude of the effect did not

change over time, as indicated by nonsignificant Performance-Avoidance Goal  $\times$  Time interaction term.

#### *Do Changes in Achievement Goals Have Implications for Changes in Achievement?*

We conducted growth-curve analyses with two different types of predictors of achievement goals to examine the effects of intraindividual changes in achievement goals.<sup>8</sup> In line with procedures outlined by Singer and Willett (2003), we used Time 1 centering to create two achievement-goal predictors, one representing the initial level of goals and one representing gains and losses from the initial level. The initial levels of goals are Level 2 between-individual variables (termed baseline score,  $Time_1 Goal = Goal_{1j}$ ), and deviation scores from the initial levels are Level 1 within-individual predictors of achievement goals (termed  $DevT_1$  scores,  $DevT_1 Goal = Goal_{ij} - Goal_{1j}$ ).  $DevT_1$  scores were computed by subtracting the Time 1 goal scores from goal scores at four time points, and thus, they represent the deviation at each time point from the initial baseline. The significant effects of  $DevT_1$  scores indicate that the students' goal changes from their own initial baseline affect their growth trajectories of GPA. The

<sup>7</sup> Interactions between goals, gender, and race were tested. None of these interactions were significant, and hence, these interaction terms were dropped from the final models.

<sup>8</sup> Analyses with Time 1-centered variables are statistically equivalent with those with raw goal scores (time varying) but enable a different interpretation (see Singer & Willett, 2003, for discussion of this approach).

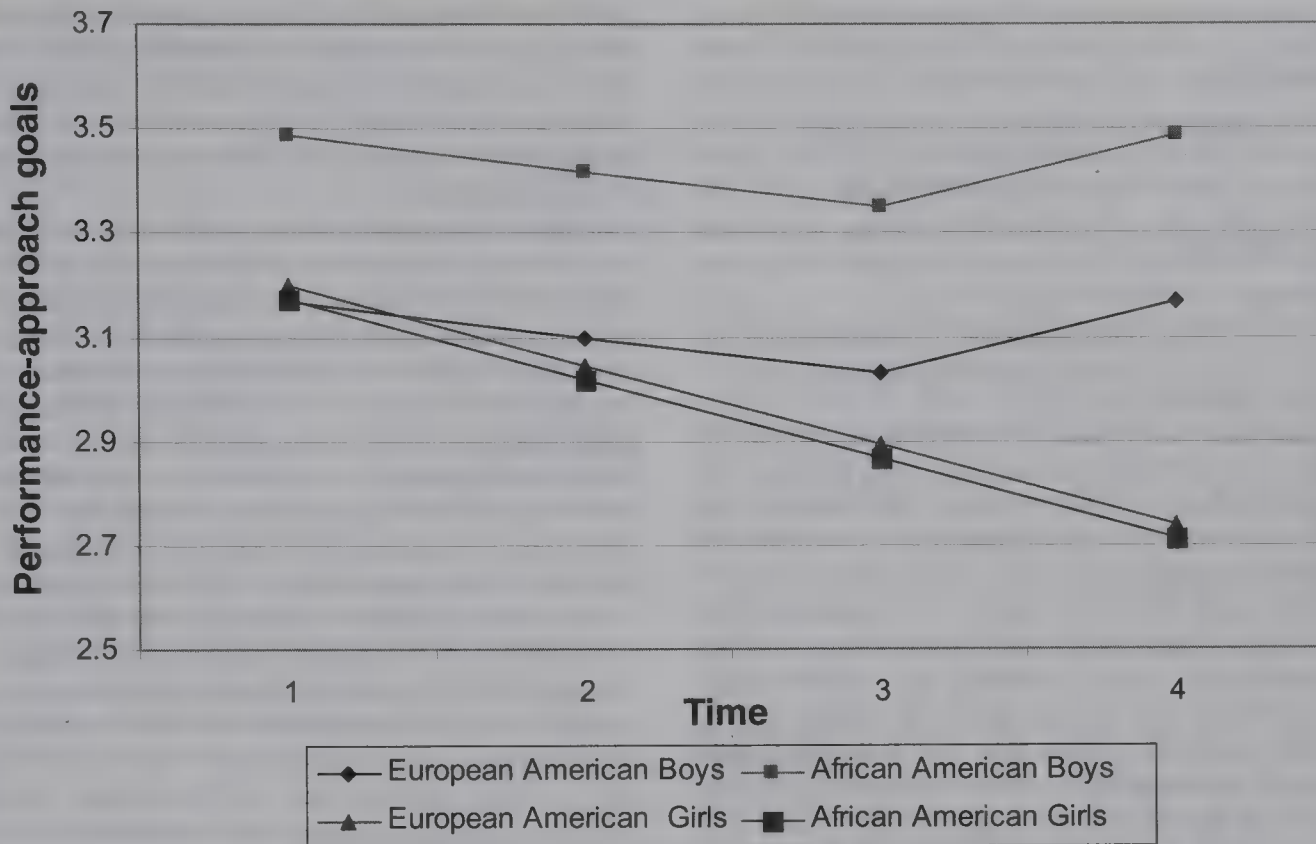


Figure 2. Gender and race differences in the development of performance-approach goals.

HLM equation that estimates the effects of achievement goals on growth trajectories of achievement with Time 1 centering method takes the following form:

$$\begin{aligned} \text{Level 1: } \text{GPA}_{ij} = & \beta_{0j} + \beta_{1j}(\text{Time})_{ij} + \beta_{2j}(\text{DevT}_1\text{MG})_{ij} \\ & + \beta_{3j}(\text{DevT}_1\text{PAP})_{ij} + \beta_{4j}(\text{DevT}_1\text{PAV})_{ij} \\ & + \beta_{5j}(\text{Time})_{ij}(\text{DevT}_1\text{MG})_{ij} + \beta_{6j}(\text{Time})_{ij}(\text{DevT}_1\text{PAP})_{ij} \end{aligned}$$

$$+ \beta_{7j}(\text{Time})_{ij}(\text{DevT}_1\text{PAV})_{ij} + \epsilon_{ij},$$

where  $(\text{DevT}_1\text{MG})_{ij}$ ,  $(\text{DevT}_1\text{PAP})_{ij}$ , and  $(\text{DevT}_1\text{PAV})_{ij}$ , represent deviation scores from Time 1 scores for mastery, performance-approach, and performance-avoidance goals, respectively.

$$\text{Level 2: } \beta_{0j} = \gamma_{00} + \gamma_{01}(\text{Gender})_j + \gamma_{02}(\text{Race})_j + \gamma_{03}(\text{Gender})_j$$

Table 5  
The Association Between Achievement Goals and GPA at Time 1

Model parameter	Estimate (95% confidence interval)	SE	df	t
Model for initial status				
M initial status	3.37 (2.65, 4.10)	.37	136	9.19***
Fifth-grade GPA	0.68 (.62, .73)	.03	77	25.94***
Race	0.81 (.54, 1.08)	.14	1230	5.81***
Model for time-varying predictors				
Mastery goals	0.00 (-.14, .13)	.07	193	-0.06
Performance-approach goals	-0.19 (-.33, -.05)	.07	31	-2.77**
Performance-avoidance goals	-0.09 (-.16, -.02)	.04	224	-2.43*
Model for linear change				
M change rate	-1.33 (-1.62, -1.05)	.15	573	-9.19***
Model for time by time-varying predictors				
Time × Mastery Goals	0.13 (.07, .20)	.04	396	3.82***
Time × Performance-Approach Goals	0.08 (.02, .14)	.03	824	2.75**

Note. GPA = grade point average. The omitted category is African Americans. Estimates for European Americans are shown in the table. *dfs* are from PROC MIANALYZE (SAS, 2003), not from hierarchical linear modeling.

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .



$$\begin{aligned} & \times \text{Race})_j + \gamma_{04}(\text{Grade5 GPA})_j + \gamma_{05}(\text{Time}_1\text{MG})_j \\ & + \gamma_{06}(\text{Time}_1\text{PAP})_j + \gamma_{07}(\text{Time}_1\text{PAV}) + U_{0j} \\ \beta_{1j} = & \gamma_{10} + \gamma_{11}(\text{Gender})_j + \gamma_{12}(\text{Race})_j + \gamma_{13}(\text{Gender} \\ & \times \text{Race})_j + \gamma_{14}(\text{Time}_1\text{MG})_j + \gamma_{15}(\text{Time}_1\text{PAP})_j \\ & + \gamma_{16}(\text{Time}_1\text{PAV})_j + U_{1j} \end{aligned}$$

For the remaining  $k = 2$  through 7,  $\beta_{kj} = \gamma_{k0}$ , where  $(\text{Time}_1\text{MG})_j$ ,  $(\text{Time}_1\text{PAP})_j$ , and  $(\text{Time}_1\text{PAV})_j$  represent Time 1 scores for mastery, performance-approach, and performance-avoidance goals, respectively.

Because of the inclusion of the time varying predictors at Level 1, the intercept ( $\gamma_{00}$ ) and the slope parameter ( $\gamma_{10}$ ) represent the average initial level and the rate of change after adjusting for fifth-grade GPA, gender and racial differences, and goal effects. If both the main effects of goals (i.e.,  $\gamma_{20}$ ,  $\gamma_{30}$ ,  $\gamma_{40}$ ) and the Goal  $\times$  Time interaction terms (i.e.,  $\gamma_{50}$ ,  $\gamma_{60}$ ,  $\gamma_{70}$ ) are significant, the results would indicate that achievement goals predict both the initial levels and the rates of changes. If only the main effect term of a goal is significant, then the goal affects the level of growth trajectory and the effects are constant over time. If the main effect of goal and the effect of interaction term are in different signs, then the results indicate that the effects of goal undergo qualitative change over time. Insignificant terms are dropped so that final models include only significant fixed effects at the .05 level. The results are shown in Table 6.

### Predicting GPA

**Mastery goals.** Baseline mastery goal scores did not predict the initial levels; however, they positively influenced the rates of

change in GPA over time. Thus, students who initially had high mastery goal scores showed more desirable patterns of changes in GPA over time. In addition, the  $\text{Dev}T_1$  positively predicted the growth trajectories of GPA. Thus, gains from the initial baseline score of mastery goals yielded additional benefits in terms of rate of change in GPA.

**Performance-approach goals.** Baseline performance-approach goal scores predicted neither the levels nor the rates of changes; however, the  $\text{Dev}T_1$  performance-approach goal scores had a negative effect on the intercept but a positive effect on the slope. Thus, inclines or declines from the baseline affected the developmental trajectories of GPA, but as we found in the previous analysis, the effect varied over time. That is, students who showed increases in performance-approach goals from their initial baseline showed more adaptive patterns of changes in GPA, but because of the negative effect on the initial level of GPA, the positive effect on the slope did not compensate for the initial negative effect.

**Performance-avoidance goals.** Baseline performance-avoidance goal scores predicted initial levels but not the rate of change of GPA. The results indicate that baseline performance-avoidance goals predicted consistently lower levels of GPA across all time points. Gains from the initial baseline lowered the levels of GPA even more, as indicated by the significant effects of  $\text{Dev}T_1$  performance-avoidance goal scores on the initial levels of GPA. Taken together, the results supported the conclusion that performance-avoidance goals were associated with consistently lower levels of GPA.

### Supplemental Analyses to Address Causality

In the current study, we examined whether achievement goals predict achievement. The temporal sequence of our design (our

Table 6  
The Effects of Achievement Goals on Growth Trajectories of GPA, Using Time 1 Centering Technique

Model parameter	Estimate (95% confidence interval)	SE	df	t
Initial level				
M initial status	3.04 (2.11, 3.96)	.47	222	6.48***
European American boys	0.67 (.31, 1.02)	.18	1963	3.70***
African American boys	-0.25 (-.63, .14)	.20	580	-1.26
European American girls	0.85 (.53, 1.17)	.17	12749	5.14***
Fifth-grade GPA	0.67 (.61, .72)	.03	52	24.47***
T1 mastery goal	0.04 (-.13, .21)	.09	488	0.47
T1 performance-approach goal	-0.05 (-.24, .13)	.09	78	-0.59
T1 performance-avoidance goal	-0.14 (-.27, -.01)	.07	338	2.14*
Time-varying predictors				
$\text{Dev}T_1$ mastery goal	0.00 (-.27, .26)	.13	19	-0.02
$\text{Dev}T_1$ performance-approach goal	-0.34 (-.55, -.13)	.10	26	-3.29**
$\text{Dev}T_1$ performance-avoidance goal	-0.07 (-.15, .01)	.04	434	-1.83†
Linear change				
M rate of change	-1.30 (-1.68, -.93)	.19	679	-6.84***
T1 mastery goal	0.16 (.06, .26)	.05	57	3.21**
T1 performance-approach goal	0.05 (-.03, .13)	.04	137	1.21
$\text{Dev}T_1$ mastery goal	0.12 (.01, .22)	.05	57	2.28*
$\text{Dev}T_1$ performance-approach goal	0.15 (.04, .26)	.05	22	2.92**

Note. GPA = grade point average; T1 = Time 1;  $\text{Dev}T_1$  = deviation from T1 score. The omitted category is African American girls. Thus, the estimate for mean initial status represents the mean initial level for African American girls. The estimates for other groups indicate the difference from African American girls. *dfs* are from PROC MIANALYZE (SAS, 2003), not from hierarchical linear modeling.

† $p < .10$ . \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

measurements of achievement goals are at the beginning of the semester, and GPA represents achievement across the entire semester), as well as controlling for prior achievement, helped address reciprocal causation concerns but did not preclude the possibility that achievement influences subsequent achievement goals. Thus, we conducted HLM analyses to estimate the effect of prior achievement on growth trajectories of achievement goals (see Table 7 for the results). Previous semester GPA from spring of fifth grade to fall of seventh grade were entered as a time-varying predictor for growth trajectories of achievement goals from fall of sixth grade to spring of seventh grade. Prior achievement level did not predict the growth trajectories of mastery goals. Prior achievement level predicted the levels but not the rates of change in the growth trajectories of performance-approach goals and avoidance goals. These results indicated that roughly a one-letter-grade difference in GPA yielded a one-point drop in performance-approach and avoidance goals in the following semesters. Students who did well in the previous semester were less likely to endorse either performance-approach or avoidance goals. These relationships between prior achievement and subsequent goal adoption did not change over time, as indicated by nonsignificant Goal  $\times$  Time interaction. Results are provided in Table 7.

### Discussion

Motivation, learning, and achievement are not static. However, most of the research on a prominent motivational construct, the achievement goal, has considered measurements of goals at only one or two time points. Thus, the research design and analytic methods have not captured the changing nature of achievement

goals. With a longitudinal design and growth-curve analytic techniques that incorporated both measurements of achievement goals and achievement across time, the present study expanded current understanding of the nature and consequences of achievement goals. Our developmental approach provided important insights into the development of achievement goals, the association of achievement goals with achievement across time, and the consequences of changes in achievement goals for achievement.

### The Development of Achievement Goals

Consistent with prior research reporting maladaptive changes in motivation, engagement, and achievement during early adolescence (Eccles et al., 1997), achievement goals were found to decline in general, regardless of the type. Previous longitudinal research has assessed achievement goals at only two points in time and in different grades. With four measurements of achievement goals (fall and spring of sixth grade and fall and spring of seventh grade), we were able to distinguish within-school-year from between-school-year changes. The general pattern across these four time points was a decline from fall to spring in sixth grade, stability from spring of sixth grade to fall of seventh grade, and a decline from fall of seventh grade to spring of seventh grade. Previous research has implicated the transition to middle school as the source of the decline in goals (E. M. Anderman & Midgley, 1997; L. H. Anderman & Anderman, 1999). However, with more time points, we found that the major source of the overall decline was within year, not between years. The average level of achievement goals in the beginning of middle school was similar to that in the spring of elementary school, which suggests that moving into

Table 7  
*The Effect of Prior Achievement Level on Adoption of Achievement Goals*

Model parameter	Coefficient (95% confidence interval)	SE	df	t
Mastery goals				
Initial status				
M initial status	3.57 (3.42, 3.71)	.07	270	48.39***
Time-varying predictors				
Previous semester GPA	0.00 (−.02, .02)	.01	112	−0.02
Linear change				
M change rate	−0.09 (−.12, −.06)	.02	130	−5.74***
Performance-approach goals				
Initial status				
M initial status	3.51 (3.35, 3.67)	.08	5354	43.47***
Time-varying predictors				
Previous semester GPA	−0.03 (−.05, −.01)	.01	12020	3.64***
Linear change				
M change rate	−0.12 (−.16, −.09)	.02	481	−7.81***
Performance-avoidance goals				
Initial status				
M initial status	3.55 (3.34, 3.75)	.11	125	33.61***
Time-varying predictors				
Previous semester GPA	−0.04 (−.06, −.02)	.01	168	−3.75***
Linear change				
M change rate	−0.17 (−.21, −.13)	.02	165	−8.72***

Note. GPA = grade point average. *dfs* are from PROC MIANALYZE (SAS, 2003), not from hierarchical linear modeling.

\*\*\**p* < .001.



a new, larger school environment was not immediately a catalyst for dramatic shifts in level of goals.

A study concerning the change in elementary-school students' achievement goals for different types of literary tasks found declines from fall to spring (Meece & Miller, 2001). Thus, the dissipation of goals in a similar classroom setting may be a trend that is in place prior to middle school. The finding that declines in goals occur with more experience in a specific classroom merits more attention. What aspects of the classroom environment are contributing to this? Does novelty, or an element of the unknown, heighten the pursuit of goals? Much attention has been devoted to classroom features that are associated with student goal pursuit, but researchers generally focus on between-class differences of data at one time point (Wigfield et al., 2006). Future research that examines how classroom features are related to the development of goals within the same environment could provide important insights into how teachers might best support students' goals.

For mastery and performance-avoidance goals, the developmental pattern was the same for all groups. However, for performance-approach goals, gender and race were important in understanding the developmental patterns (see Figure 3). For girls, performance-approach goals showed a similar pattern to that of the other goals (within-year declines for both elementary- and middle-school years). For boys, performance-approach goals declined during elementary school but increased during middle school. Considered in tandem with a Gender  $\times$  Race interaction on the intercept, the results indicated that African American boys had the highest level of performance-approach goals of all students, at all time points, whereas European American boys were similar to girls in elementary school but had higher levels of performance-approach goals than did girls by the end of the first year in middle school. It is interesting that boys' performance-approach goals increase after

moving into a context that is typically described as more competitive and performance oriented (Eccles et al., 1997; Midgley, 1993). In general, boys' social interactions with peers involve more competition and dominance seeking than do girls' (e.g., Hartup, 1989; Maccoby, 1990; Rose & Rudolph, 2006). Perhaps there is a match between this gendered characteristic and the middle-school context that causes performance-approach goals to increase for boys and not girls. Being better than others in the classroom may become more important to boys during early adolescence because it is in line with societal ideals of masculinity that emphasize competition, winning, and dominance. These issues may be salient for African American boys even earlier, as they had higher performance-approach goals than all other students, even before the increase for boys that occurred in middle school. It is concerning that African American boys had higher performance-approach goals than did all other students in elementary school, given that performance-approach goals had a deleterious impact on achievement in that setting.

Although the developmental patterns for mastery and performance-avoidance goals were the same for all groups, there were gender and race differences in the levels across time. Consistent with prior research, girls and African American students were higher in mastery goals. Also consistent with prior research, boys were higher in a performance-avoidance goal than were girls. However, there were no interactive effects of either gender or race with goals on achievement, which indicated that the relation of goals to achievement was the same for all groups across time. Thus, similar to other aspects of motivation (e.g., ability perceptions), the mean levels may vary, but the nature of the consequences do not (see Graham, 1994, for a review). Future research could further our understanding of group differences by examining if the antecedents of achievement goals are similar for all students.

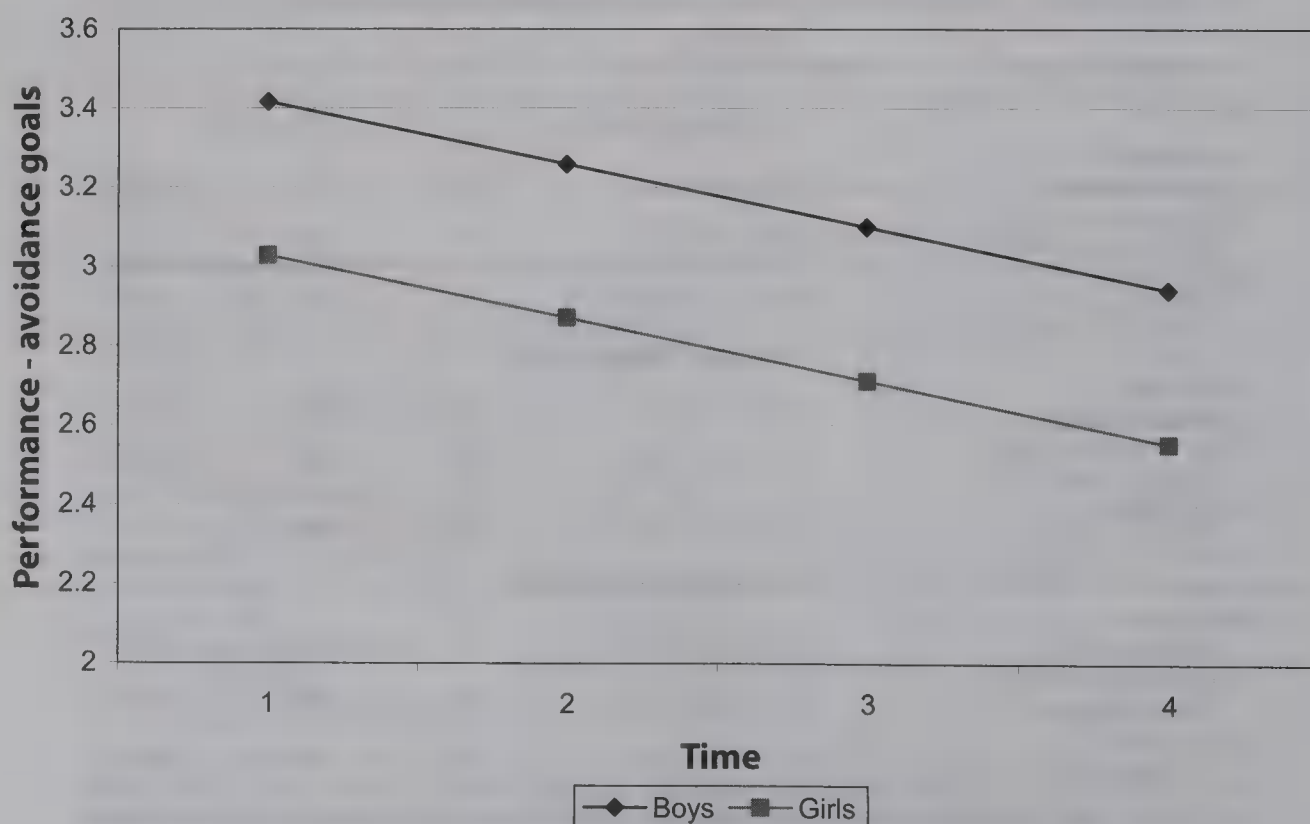


Figure 3. Gender differences in the development of performance-avoidance goals.

It may be that gender and/or ethnicity play a role in students' perceptions of or experiences in classrooms that may explain the different levels of achievement goals.

### *The Changing Association of Achievement Goals and Achievement Across Time*

By treating achievement goals as time-varying predictors of achievement across time, we were able to examine whether the associations of achievement goals to achievement change during early adolescence. Our sample made a transition into middle school during this time, which allowed us to examine these associations across contexts that vary in terms of emphasis on competition and social comparison (see Eccles et al., 1997; Juvonen et al., 2004; Midgley, 1993). When we controlled for prior achievement, a performance-avoidance goal had a consistent negative association with subsequent achievement across the four time points, which is in line with prior research that consistently documents the negative consequences of such goals. When we controlled for prior achievement, mastery goals were positive predictors of subsequent achievement at Times 2–4. Thus, mastery goals fostered the achievement of young adolescents, especially in the middle-school context. Findings for a performance-approach goal captured changing associations across time. A performance-approach goal was detrimental to achievement in the elementary-school context, particularly at Time 1 (negative association was significant at Time 1 and marginally significant at Time 2) and then became unrelated to achievement when students moved into middle school.

It has been suggested that as students advance through the educational system, the context becomes more competitive, and the consequences of achievement goals may change; a performance-approach goal may become more adaptive, and a mastery goal may become less adaptive for achievement (Midgley et al., 2001). The findings for a performance-approach goal were somewhat in line with this idea, as the detrimental effects on achievement dissipated as students moved into middle school. The findings for a mastery goal were contrary to this idea, as mastery goals became increasingly positive predictors of achievement as students moved into middle school. One possibility is that a mastery goal is especially facilitative of achievement in challenging academic situations. Grant and Dweck (2003) found that a mastery goal predicted more adaptive coping and higher achievement in the face of challenge for undergraduate students enrolled in a difficult premedical course. Students in our sample navigated a transition from a small elementary school to a large middle school. Against this backdrop, we saw GPA decline from elementary school to middle school. Thus, students likely experienced academic challenge or uncertainty during this transition. Our data indicate that a mastery goal supported students' achievement during this transition. However, a mastery goal was a positive predictor of achievement in the spring of elementary school, so this would not be a complete explanation.

The pattern of results for achievement goals and achievement are in sharp contrast to those consistently found for college students. For college students, the relationship between achievement goals and achievement is consistently null for a mastery goal and positive for a performance-approach goal (see Harackiewicz, Barron, Pintrich, et al., 2002, for a review). However, our findings are

in line with much research on younger samples (e.g., Kaplan & Maehr, 1999; Pajares & Valiante, 2001; A. M. Ryan et al., 2005; Wentzel, 1993). Thus, conclusions about which achievement goals are best for learning and achievement are different for younger students in school and older adolescents in college. In addition to the educational climate becoming more competitive as students progress through the system, there are numerous other differences in context, sample, and developmental stage that may contribute to an explanation of the different patterns. Young adolescent students learn among their peers in classes all day, in contrast with college students who direct their own learning for much of the day. Young adolescent students have frequent assignments and feedback on their learning, whereas college students may only have two or three papers or exams per semester. Younger samples in public schools represent heterogeneity of ability levels from the broad spectrum of society, whereas samples taken from college represent only the upper ability levels. All of these differences may contribute to an explanation of why mastery goals are important to the achievement of young adolescents but not to college students.

The implication of our research is that a mastery goal should be encouraged in early adolescent students because it supports learning and achievement. In recent years (and primarily on the basis of research on college students), mastery goals have been lauded as beneficial for supporting intrinsic motivation but not achievement (e.g., Harackiewicz, Barron, Pintrich, et al., 2002; Harackiewicz & Tauer, 2006). However, achievement is often the main concern of administrators and educators, especially given the current climate of standardized testing brought on by the No Child Left Behind Act of 2001 (2002). If mastery goals are not seen as having benefits for achievement, it is unlikely they will garner much attention for school reform. This could represent a missed opportunity for children in our schools, as much work has been done regarding how to create an educational climate that encourages mastery goals (e.g., Ames, 1992; Maehr & Midgley, 1996). Thus, it is critical to take a developmental perspective and distinguish patterns for younger students from patterns found for college students regarding achievement goals and achievement.

### *The Implications of Changes in Achievement Goals for Changes in Achievement*

The benefits of mastery goals for achievement were also seen in our third set of analyses, in which we used a within-individual approach to assess the development of students' achievement goals. Our first two sets of analyses concerned general patterns and variations between students (the first pertaining to the development of goals and the second pertaining to the associations of goals to achievement across time). Although general patterns are important in understanding normative shifts that occur, they do not address the significant variability in these patterns that exists for individual students. Regarding the development of goals, not all students conform to the general declining pattern of achievement goals. Regarding the changing goals–achievement relationship, there are intraindividual differences that tell us something different about the motivation–achievement link than do the interindividual differences. Similar to most prior research, our examination of general patterns of the link between achievement goals and achievement across time concerned how high or low scores in different types of goals, compared with other students, explained



subsequent variations in achievement. In our final set of analyses, we focused on within-individual change to determine if inclines or declines from student's initial level of goals were important to understanding students' personal trajectory of achievement. Focusing on intraindividual change captures the dynamic nature of achievement goals that is widely acknowledged in theory but seldom captured in analytic approaches.

Fluctuations in mastery and performance-approach goals within individual students were important for a thorough understanding of the effects of achievement goals. In particular, deviation scores provided insight into how a mastery goal promotes positive changes in achievement across time. Above and beyond baseline scores, changes in mastery goals promoted changes in achievement. When we controlled for initial level, students who increased their mastery goals exhibited greater gains in achievement across time. Documenting such effects has important educational implications because it indicates that even for students with a "maladaptive" goal profile (i.e., low mastery goals), teachers only need to move them in the right direction for students to reap some achievement benefits. Encouraging a student to intensify his or her mastery goal may help students actually master the material. Personal reminders from teachers about the importance of a mastery goal may sustain or boost students' focus on mastery goals throughout the year, and such positive developments will, in turn, sustain or improve their achievement over time.

### *Prior Achievement as a Predictor of Achievement Goals*

Our supplemental analyses indicated that prior achievement was not related to subsequent mastery goals and was negatively related to performance goals (both approach and avoidance). When students' grades increased, their performance goals were more likely to go down. These findings are not what one would expect on the basis of prior theory and research indicating that high perceptions of competence lead to approach goals and low perceptions of competence lead to avoidance goals (Elliot, 2005). Because perceived competence and academic performance are moderately to highly correlated, one would expect a similar pattern for GPA. However, the nature of our design and analyses make them hard to directly compare with previous research. We controlled for prior achievement and modeled intraindividual changes in GPA. Thus, we examined the effects of changes in GPA, controlling for prior levels, on changes in achievement goals. With this approach, we documented that changes in achievement were important for performance goals and not for mastery goals. Our results suggest that changes in achievement may affect students' adoption of mastery versus performance goals, rather than approach versus avoidance goals.

It is encouraging that prior achievement is not important to mastery-goal pursuit, as it suggests that low achievers may be just as amenable to intervention as high achievers. We know a great deal about teacher behaviors that are important to students' goals. Ames (1990, 1992) has described how teacher practices in six categories (Task, Authority, Recognition, Grouping, Evaluation, Time) can promote a mastery-goal emphasis in classrooms. Qualitative work has expanded on Ames's work and noted how affect and teacher-student relationships are also critical to fostering a mastery-oriented climate (Patrick, Anderman, Ryan, Edelin, & Midgley, 2001; Patrick, Turner,

Meyer, & Midgley, 2003; Turner et al., 2002). Linnenbrink (2005) designed an intervention based on feedback to small groups that was successful in increasing students' mastery goals. Blackwell, Trzesniewski, and Dweck (2007) designed a Web-based intervention that increased students' malleable theory of intelligence, which in turn increased students' mastery goals and grades. Thus, although there are challenges in the current educational climate (see Urdu & Turner, 2005), there is much knowledge about how to encourage mastery goals in students. All of this research could be drawn on to promote mastery goals in early adolescent students and to best support their learning and achievement.

### *Limitations and Future Research*

Although our developmental approach provided new insights regarding achievement goals, there are several limitations that need to be considered and possibly addressed in future research. First, our sample only represented two ethnic groups, and it is therefore not known whether the results would generalize to other ethnic groups. Second, we only examined grades as indicators of achievement. This was appropriate given our focus on within-year and between-year development of goals and achievement. Nonetheless, standardized tests provide different information about students' learning and achievement than do grades, and future work that examines the link between achievement goals and standardized tests across time could broaden our understanding of the implications of goals for achievement (see K. Ryan & Ryan, 2005; K. Ryan, Ryan, Arbutnot, & Samuels, 2007, for a discussion of achievement goals and standardized-test performance).

Third, a longer time frame would have provided additional understanding of the development of achievement goals. With our four time points of data, we documented within-year, but not between-year, declines in achievement goals in both sixth and seventh grades, indicating that the transition did not immediately have an impact on the average level of goals. However, it may be that the typical yearly pattern for students' goals is to decline across the year and then recover over the summer break, beginning high again the following year. It is possible that the middle-school transition does have an impact on the average level of student goal pursuit by disrupting this cycle. Future research encompassing additional years before the transition could further our understanding of typical cycles of motivation. In a cross-sectional study, Pajares and Cheong (2003) found nonlinear changes in mastery goals and performance-approach goals (decrease from elementary school to middle school and increase in high school for mastery goals and decrease from elementary school to middle school and no change in high school for performance-approach goals). There was no change in performance-avoidance goals. This is another intriguing developmental pattern that could be further investigated with longitudinal data. Although practically challenging, longitudinal studies following students from elementary school to college could better examine the proposition that performance-approach goals exert a positive effect on achievement and mastery goals become null as students advance through the educational system.

Fourth, we asked students about their achievement goals for schoolwork in general, and our results may not generalize across



all subjects. Recent research on achievement goals has investigated students' general goals as well as subject-specific goals.<sup>9</sup> Both approaches are valid because achievement goals are conceptualized as a function of individual differences (i.e., personality, dispositions, or general motives) as well as due to features of the immediate environment. Students tend to approach schoolwork in different subjects in a similar way, as indicated by moderate to large correlations between the goals in different subjects (Bong, 2001; Elliot & McGregor, 2001). Further, achievement goals operate similarly across different academic subject areas (Bong, 2001; Wolters et al., 1996), so measuring at a general level is unlikely to confound results. Nonetheless, there are differences in the nature of knowledge and teaching between subjects that may contribute to different developmental patterns at the subject-specific level, and this could be a fruitful area for future research.

## Conclusion

By taking advantage of current advances in growth-curve analytic techniques, we were able to garner a more complete understanding of students' achievement goals and the importance of these goals for achievement. Our focus on the development of goals across time in early adolescence revealed important patterns for this age group. At this stage of life, goals exhibit within-year rather than between-year decreases (with the exception of boys' performance-approach goals, which increase in middle school). As students move through early adolescence, mastery goals become increasingly beneficial for achievement. Performance-approach goals are detrimental in elementary school and provide no benefits in middle school. Performance-avoidance goals consistently undermine achievement throughout early adolescence. Personal increases in mastery goals over time provide a boost to student achievement. The overall implications of the present research is that mastery goals should be supported and encouraged in early adolescent students to best support their achievement. This recommendation is a deviation from recent conclusions based on data from college students regarding achievement goals and achievement. Thus, results highlight that a developmental perspective is critical to understanding achievement goals and achievement. More generally, we hope the results highlight the contribution a developmental approach can make toward understanding motivation and achievement. It is only with a complete understanding of the development of motivation, engagement, and achievement in school that we can make recommendations to educators and parents about what matters for student learning.

<sup>9</sup> Our review of articles concerning achievement goals in the last 5 years of the *Journal of Educational Psychology* indicated that about half of the research measured goals at the general level and about half measured goals specific to a certain subject or class.

## References

Ablard, K. E., & Lipschultz, R. E. (1998). Self-regulated learning in high-achieving students: Relations to advanced reasoning, achievement goals, and gender. *Journal of Educational Psychology, 90*, 94–101.

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage.

Ames, C. (1992). Classrooms: Goals, structures, and student motivation. *Journal of Educational Psychology, 84*, 261–271.

Ames, C. A. (1990). Motivation: What teachers need to know. *Teachers College Record, 91*, 409–421.

Ames, C., & Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation processes. *Journal of Educational Psychology, 80*, 260–267.

Anderman, E. M., & Midgley, C. (1997). Changes in achievement goal orientations, perceived academic competence, and grades across the transition to middle-level schools. *Contemporary Educational Psychology, 22*(3), 269–298.

Anderman, L. H., & Anderman, E. M. (1999). Social predictors of changes in students' achievement goal orientations. *Contemporary Educational Psychology, 24*(1), 21–37.

Blackwell, L. S., Trzesniewski, K., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development, 78*(1), 246–263.

Bong, M. (2001). Between and within-domain relations of academic motivation among middle and high school students: Self-efficacy, task-value, and achievement goals. *Journal of Educational Psychology, 93*, 23–34.

Brophy, J. (2005). Goal theorists should move on from performance goals. *Educational Psychologist, 40*(3), 167–176.

Butler, R. (1993). Effects of task- and ego-achievement goals on information seeking during task engagement. *Journal of Personality and Social Psychology, 65*(1), 18–31.

Dearing, E., Kreider, H., Simpkins, S., & Weiss, H. B. (2006). Family involvement in school and low-income children's literacy performance: Longitudinal associations between and within families. *Journal of Educational Psychology, 98*, 653–664.

Dweck, C. S., & Leggett, E. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95*, 256–273.

Eccles, J., (2005). Schools, academic motivation and stage-environment fit. In R. M. Lerner & L. Steinberg (Eds.), *Handbook of adolescent psychology* (pp. 125–154). Hoboken, NJ: Wiley.

Eccles, J. S., Lord, S. E., Roeser, R. W., Barber, B. L., & Jozefowicz, D. M. (1997). The association of school transitions in early adolescence with developmental trajectories through high school. In J. Schulenberg, J. L. Maggs, & K. Hurrelmann (Eds.), *Health risks and developmental transitions during adolescence* (pp. 283–320). New York: Cambridge University Press.

Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist, 34*, 149–169.

Elliot, A. J. (2005). A conceptual history of the achievement goal construct. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 52–72). New York: Guilford Press.

Elliot, A. J., & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology, 72*, 218–232.

Elliot, A. J., & Harackiewicz, J. M. (1996). Approach and avoidance achievement goals and intrinsic motivation: A mediational analysis. *Journal of Personality and Social Psychology, 70*, 461–475.

Elliot, A. J., & McGregor, H. A. (2001). A 2 × 2 achievement goal framework. *Journal of Personality and Social Psychology, 80*, 501–519.

Elliot, A. J., McGregor, H. A., & Gable, S. (1999). Achievement goals, study strategies, and exam performance: A mediational analysis. *Journal of Educational Psychology, 91*, 549–563.

Freeman, T., & Anderman, L. H. (2005). Changes in mastery goals in urban and rural middle school students. *Journal of Research in Rural Education, 20*(1). Retrieved August 2005 from <http://www.umaine.edu/jrre/20-1.htm>

Gehlbach, H. (2006). How changes in students' goal orientations relate to outcomes in social studies. *Journal of Educational Research, 99*, 358–370.



- Graham, S. (1994). Motivation in African Americans. *Review of Educational Research*, 64(1), 55–117.
- Grant, H., & Dweck, C. S. (2003). Clarifying achievement goals and their impact. *Journal of Personality and Social Psychology*, 85, 541–553.
- Harackiewicz, J. M., Barron, K. E., & Elliot, A. J. (1998). Rethinking achievement goals: When are they adaptive for college students and why? *Educational Psychologist*, 33(1), 1–21.
- Harackiewicz, J. M., Barron, K. E., Pintrich, P. R., Elliot, A. J., & Thrash, T. M. (2002). Revision of achievement goal theory: Necessary and illuminating. *Journal of Educational Psychology*, 94, 638–645.
- Harackiewicz, J. M., Barron, K. E., Tauer, J. M., & Elliot, A. J. (2002). Predicting success in college: A longitudinal study of achievement goals and ability measures as predictors of interest and performance from freshman year through graduation. *Journal of Educational Psychology*, 94, 562–575.
- Harackiewicz, J. M., & Elliot, A. J. (1993). Achievement goals and intrinsic motivation. *Journal of Personality and Social Psychology*, 65, 904–915.
- Harackiewicz, J. M., & Tauer, J. M. (2006). From bicycle racing to school: Competition, multiple goals and multiple indicators of success in education. In Paul A. M. Van Lange (Ed), *Bridging social psychology: Benefits of transdisciplinary approaches* (pp. 239–244). Mahwah, NJ: Erlbaum.
- Hartup, W. W. (1989). Social relationships and their developmental significance. *American Psychologist*, 44, 120–126.
- Juvonen, J., Le, V., Kaganoff, T., Augustine, C., & Constant, L. (2004). *Focus on the wonder years: challenges facing the American middle school*. Santa Monica, CA: RAND Corporation.
- Kaplan, A., & Maehr, M. L. (1999). Achievement goals and student well-being. *Contemporary Educational Psychology*, 24(4), 330–358.
- Kenney-Benson, G. A., Pomerantz, E. M., Ryan, A. M., & Patrick, H. (2005). Sex differences in math performance: The role of children's approach to schoolwork. *Developmental Psychology*, 42, 11–26.
- Linnenbrink, E. A. (2005). The dilemma of performance approach goals: The use of multiple goal contexts to promote students' motivation and learning. *Journal of Educational Psychology*, 97, 197–213.
- Maccoby, E. E. (1990). Gender and relationships: A developmental account. *American Psychologist*, 45, 513–520.
- Maehr, M. L., & Midgley, C. (1996). *Transforming school cultures*. Boulder, CO: Westview Press.
- McGregor, H. A., & Elliot, A. J. (2002). Achievement goals as predictors of achievement-relevant processes prior to task engagement. *Journal of Educational Psychology*, 94, 381–395.
- Meece, J. L., & Holt, K. (1993). A pattern analysis of students' achievement goals. *Journal of Educational Psychology*, 85, 582–590.
- Meece, J. L., & Miller, S. D. (2001). Longitudinal analysis of elementary school students' achievement goals in literacy activities. *Contemporary Educational Psychology*, 26, 454–480.
- Middleton, M. J., Kaplan, A., & Midgley, C. (2004). The change in middle school students' achievement goals in math over time. *Social Psychology of Education*, 7, 289–311.
- Middleton, M. J., & Midgley, C. (1997). Avoiding the demonstration of lack of ability: An underexplored aspect of goal theory. *Journal of Educational Psychology*, 89, 710–718.
- Middleton, M. J., & Midgley, C. (2002). Beyond motivation: Middle school students' perceptions of press for understanding in math. *Contemporary Educational Psychology*, 27(3), 373–391.
- Midgley, C. (1993). Motivation and middle level schools. In P. R. Pintrich and M. L. Maehr (Eds.), *Advances in motivation and achievement: Motivation and adolescent development* (Vol. 8, pp. 191–217). Greenwich, CT: JAI Press.
- Midgley, C., Kaplan, A., & Middleton, M. (2001). Performance approach goals: Good for what, for whom, under what circumstances, and at what cost? *Journal of Educational Psychology*, 93, 77–86.
- Midgley, C., Maehr, M. L., Hicks, L. H., Roeser, R., Urdan, T., Anderman, E. M., et al. (1997). *Patterns of adaptive learning survey (PALS)*. Ann Arbor: University of Michigan.
- Midgley, C., & Urdan, T. (1995). Predictors of middle school students' use of self-handicapping strategies. *Journal of Early Adolescence*, 15, 389–411.
- Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods*, 6, 328–362.
- Newman, R. S., & Schwanger, M. T. (1995). Students' help seeking during problem solving: Effects of grade, goal, and prior achievements. *American Educational Research Journal*, 32, 352–376.
- Nicholls, J. G. (1989). *The competitive ethos and democratic education*. Cambridge, MA: Harvard University Press.
- No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002).
- Nolen, S. B. (1988). Reasons for studying: Motivational orientations and study strategies. *Cognition and Instruction*, 5(4), 269–287.
- Pajares, F., Britner, S. L., & Valiante, G. (2000). Relation between achievement goals and self-beliefs of middle school students in writing and science. *Contemporary Educational Psychology*, 25(4), 406–422.
- Pajares, F., & Cheong, Y. F. (2003). Achievement goal orientations in writing: A developmental perspective. *International Journal of Educational Research*, 39, 437–455.
- Pajares, F., & Valiante, G. (2001). Gender differences in writing motivation and achievement of middle school students: A function of gender orientation? *Contemporary Educational Psychology*, 26(3), 366–381.
- Patrick, H., Anderman, L. H., & Ryan, A. M. (2002). Social motivation and the classroom social environment. In C. Midgley (Ed.), *Goals, goal structures, and patterns of adaptive learning* (pp. 85–108). Mahwah, NJ: Erlbaum.
- Patrick, H., Anderman, L. H., Ryan, A. M., Edelin, K., & Midgley, C. (2001). Teachers' communication of goal orientations in four fifth-grade classrooms. *The Elementary School Journal*, 102, 35–58.
- Patrick, H., Turner, J. C., Meyer, D. K., & Midgley, C. (2003). How teachers establish psychological environments during the first days of school: Associations with avoidance in mathematics. *Teachers College Record*, 105, 1521–1558.
- Pintrich, P. R. (2000). An achievement goal theory perspective on issues in motivation terminology, theory, and research. *Contemporary Educational Psychology*, 25(1), 92–104.
- Roeser, R. W., Midgley, C., & Urdan, T. (1996). Perceptions of the school psychological environment and early adolescents' psychological and behavioral functioning in school: The mediating role of goals and belonging. *Journal of Educational Psychology*, 88, 408–422.
- Rogosa, D. R. (1988). Myths about longitudinal research. In K. W. Schaie, R. T. Campbell, W. Meredith, & S. C. Rawlings (Eds.), *Methodological issues in aging research* (pp. 171–210). New York: Springer Publishing Company.
- Rose, A. J., & Rudolph, K. D. (2006). A review of sex differences in peer relationship processes: Potential trade-offs for the emotional and behavioral development of girls and boys. *Psychological Bulletin*, 132, 98–131.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
- Ryan, A. M., Hicks, L., & Midgley, C. (1997). Social goals, academic goals, and avoiding seeking help in the classroom. *Journal of Early Adolescence*, 17(2), 152–171.
- Ryan, A. M., Patrick, H., & Shim, S. (2005). Differential profiles of students identified by their teacher as having avoidant, appropriate, or dependent help-seeking tendencies in the classroom. *Journal of Educational Psychology*, 97, 275–285.
- Ryan, A. M., & Pintrich, P. R. (1997). "Should I ask for help?" The role

- of motivation and attitudes in adolescents' help seeking in math class. *Journal of Educational Psychology*, 89, 329–341.
- Ryan, K., & Ryan, A. M. (2005). The psychological processes underlying stereotype threat and standardized math test performance. *Educational Psychologist*, 40(1), 53–63.
- Ryan, K., Ryan, A. M., Arbuthnot, K., & Samuels, M. (2007). Students' motivation for standardized math exams. *Educational Researcher*, 36, 1–9.
- SAS. (2003). *Statistical Analysis System* (Version 9.1). Cary, NC: SAS Institute, Inc.
- Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.
- Senko, C., & Harackiewicz, J. M. (2005). Regulation of achievement goals: The role of competence feedback. *Journal of Educational Psychology*, 97, 320–336.
- Shim, S., & Ryan, A. (2005). Changes in self-efficacy, challenge avoidance, and intrinsic value in response to grades: The role of achievement goals. *Journal of Experimental Education*, 73, 333–349.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.
- Skaalvik, E. M. (1997). Self-enhancing and self-defeating ego orientation: Relations with task and avoidance orientation, achievement, self-perceptions, and anxiety. *Journal of Educational Psychology*, 89, 71–81.
- Stipek, D., & Gralinski, J. H. (1996). Children's beliefs about intelligence and school performance. *Journal of Educational Psychology*, 88, 397–407.
- Turner, J. C., Meyer, D. K., & Midgley, C. (2003). Teacher discourse and sixth graders' reported affect and achievement behaviors in two high-mastery/high-performance mathematics classrooms. *Elementary School Journal*, 103, 357–382.
- Turner, J. C., Midgley, C., Meyer, D. K., Gheen, M., Anderman, E., Kang, Y., & Patrick, H. (2002). The classroom environment and students' reports of avoidance behaviors in mathematics: A multimethod study. *Journal of Educational Psychology*, 94, 88–106.
- Urdu, T., & Midgley, C. (2003). Changes in the perceived classroom goal structure and pattern of adaptive learning during early adolescence. *Contemporary Educational Psychology*, 28(4), 524–551.
- Urdu, T., & Turner, J. C. (2005). Competence motivation in the classroom. In A. Elliot & C. Dweck (Eds.), *Handbook of competence and motivation* (pp. 297–317). New York: Guilford Press.
- Wentzel, K. R. (1993). Motivation and achievement in early adolescence: The role of multiple classroom goals. *Journal of Early Adolescence*, 13, 4–20.
- Wigfield, A., Eccles, J. S., Schiefele, U., Roeser, R. W., & Davis-Kean, P. (2006). Development of achievement motivation. In N. Eisenberg (Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (6th ed., pp. 933–1002). New York: Wiley.
- Wolters, C. A. (2004). Advancing achievement goal theory: Using goal structures and goal orientations to predict students' motivation, cognition, and achievement. *Journal of Educational Psychology*, 96, 236–250.
- Wolters, C. A., Yu, S. L., & Pintrich, P. R. (1996). The relation between goal orientation and students' motivational beliefs and self-regulated learning. *Learning and Individual Differences*, 8(3), 211–238.

Received July 3, 2007

Revision received December 27, 2007

Accepted January 10, 2008 ■



# An Exploration of Young Adolescents' Social Achievement Goals and Social Adjustment in Middle School

Allison M. Ryan  
University of Illinois

S. Serena Shim  
Northern Arizona University

Two studies investigated the proposition that social achievement goals (different orientations toward social competence) are an important aspect of young adolescents' social motivation. Study 1 ( $N = 153$  6th-grade students) established that different orientations toward developing or demonstrating social competence can be seen in young adolescents' responses to open-ended questions about their social goals and social competence. Study 2 ( $N = 217$  6th-grade students) evaluated a new survey measure of social achievement goals for young adolescents. Exploratory factor analyses indicated a 3-factor model (social development, demonstration-approach, and demonstration-avoid goals). Different social achievement goals were associated with distinct patterns of subsequent self- and teacher-reported social adjustment (prosocial, aggressive, and anxious solitary behaviors, as well as social worry, best-friend quality, and perceived popularity). Effects for social achievement goals were independent of perceived social competence and gender.

**Keywords:** motivation, peers, social adjustment

A common portrayal of young adolescents is that this is an age group desperate for peer approval. But is it only validation from peers as being "cool" or "popular" that makes an adolescent feel socially competent? Do adolescents have an inner compass for evaluating themselves, one of personal standards gleaned from their past successes and failures and oriented toward the outcome of developing positive peer relationships? Can such an inner compass be nurtured and guide young adolescents more smoothly toward adjustment in middle school? Different orientations to competence (demonstrating vs. developing competence) have been examined extensively in the academic, athletic, and work domains. However, little is known about such orientations in the social domain, and existing work has focused on adults (Elliot, Gable, & Mapes, 2006; Horst, Finney, & Barron, 2007; Ryan & Shim, 2006; but see also Erdley, Cain, Loomis, Dumas-Hines, & Dweck, 1997). Different orientations toward achieving competence are the crux of an achievement goal approach to motivation. This approach has proven to be powerful in explaining processes and outcomes for adolescents in the academic and athletic domains (Elliot & Dweck, 2005). We propose that social achievement goals are relevant to young adolescents' social strivings with peers and will advance understanding of their social adjustment in middle school.

Social adjustment in middle school is a source of increasing concern for educators, parents, and researchers. A recent review of 20 years of research, including international comparisons, depicts

middle schools in the United States as the Bermuda Triangle of education (Juvonen, Le, Kaganoff, Augustine, & Constant, 2004). Early adolescence marks the beginning of a downward trend in academic and social adjustment for many children. U.S. middle school students have the highest rate of emotional problems and most negative views of peer culture among the 11 countries included in the report (Juvonen et al., 2004). Attention to social adjustment in middle school is important in its own right, but as social and academic adjustment are interdependent, it is also critical for realization of recent academic benchmarks for this age group (e.g., No Child Left Behind mandates). Promoting positive outcomes for young adolescents, such as prosocial behavior and positive peer relationships, and diminishing negative outcomes, such as aggressive behavior and social anxiety, are critical issues facing educators. Key to finding solutions is elucidating the factors that underlie social adjustment in middle school.

## An Achievement Goal Approach

To understand social adjustment in middle school, we take an achievement goal approach, which is different from the more typical content approach to social goals. A content approach is concerned with what outcomes individuals pursue and identifies categories of goals that characterize what individuals want (Grant & Dweck, 2003; Pintrich, 2000; Wentzel, 2000). Research has been informative in documenting the various goals that children and adolescents strive for in social situations and has highlighted that individuals often have goals for affiliation, responsibility, nurturance, and intimacy (e.g., Ford, 1992; Wentzel, 2001). On the lighter side, individuals often want companionship or fun (e.g., Anderman, 1999; Jarvinen & Nicholls, 1996; Wentzel, 2001). On the darker side, individuals may desire revenge, domination, or control over others (e.g., Chung & Asher, 1996; Erdley & Asher, 1996; McAdams, 1987; Ojanen, Gronroos, & Salmivalli, 2005; Rose & Asher, 1999). Achievement goals transcend various con-

---

Allison M. Ryan, Department of Educational Psychology, University of Illinois; S. Serena Shim, Department of Educational Psychology, Northern Arizona University.

Correspondence concerning this article should be addressed to Allison M. Ryan, Department of Educational Psychology, University of Illinois, 230 Education Building, 1310 South Sixth Street, Champaign, IL 61820. E-mail: ryan2@uiuc.edu

tent or outcome goals that are salient to individuals in social settings (Grant & Dweck, 2003; Pintrich, 2000; Wentzel, 2000) and capture key distinctions in orientations to competence (Dweck, 1986; Elliot, 2005). Regardless of whether people want intimacy or fun or both in a social situation, it is likely they also want to feel socially competent. Whether they are oriented to demonstrate their social competence, develop their social competence, or possibly to do both has implications for their beliefs and behavior. An achievement goal approach does not compete with a content approach, but instead provides a new angle, and thus it will complement and expand current understanding of young adolescents' social motivation and adjustment. Because the present research builds on a theoretical framework rooted in the academic domain, it will be informative about general motivational processes that promote or hinder adaptation across domains.

In the present research we expand Ryan and Shim's (2006) conceptualization of social achievement goals (social development, social demonstration-approach, and social demonstration-avoid) to young adolescents. We examine three achievement goals that are analogous to those identified in other domains, although described with different terms (e.g., "mastery" and "performance" in the academic domain, see Elliot & McGregor, 2001; and "task" and "ego" in the sports domain, see Duda & Nicholls, 1992). A *social development goal* is concerned with developing social competence with peers. The focus is on learning new things, growth, and improvement. Success is judged by whether one is improving social skills, deepening the quality of relationships, or developing social life in general. A *social demonstration-approach goal* focuses on demonstrating social competence and gaining from peers positive judgments that one is socially desirable. A *social demonstration-avoid goal* is concerned with demonstrating that one does not lack social competence. The focus is on avoiding doing something that would incur negative judgments from peers and indicate social undesirability. With both social demonstration goals, attention focuses on the appearance of the self, especially in relation to others. For a social demonstration-approach goal, success is garnering positive feedback from peers, attaining social prestige, and having a good reputation compared to others (e.g., being popular or seen as important); for a social demonstration-avoid goal, success is avoiding negative judgments from peers compared to others and lacking a reputation as socially awkward or ineffective (e.g., not being seen as a "loser" or "geek"). Similar to the distinction made for demonstration goals, there has been some attention to bifurcating development goals into separate approach and avoid goals, although Ryan and Shim did not include this in their conceptualization of achievement goals in the social domain (neither did Horst et al., 2007, nor Elliot et al., 2006). Such development-avoid goals have been viewed as less prevalent in real-world situations and have weaker relations with adjustment variables in the academic domain (Elliot, 2005). Given the weaker conceptual and practical justifications for a development-avoid goal, we do not include it in our first investigation of social achievement goals in young adolescents.

### Overview of Present Research

The overarching goal of the present research is to examine whether social achievement goals are an important aspect of social motivation for young adolescents and have implications for their

social adjustment in school. Study 1 is a pilot study in which we examine whether social achievement goals are relevant for young adolescents' social motivation by collecting open-ended data about their descriptions of social goals and social competence. This is an important first step because, unlike most experiments or surveys, this can illustrate that different orientations toward social competence can be seen when young adolescents describe their social goals in their own life, in their own words. In Study 2, informed by Study 1, we create new survey items to assess young adolescents' social achievement goals. We explore the structure and nature of social achievement goals and investigate their implications for key aspects of social adjustment in middle school (prosocial behavior, overt aggressive behavior, anxious solitary behavior, social worry, positive features within close friendships, and perceived popularity). Recent research has examined social achievement goals in adults and found that they are distinct factors that differentially relate to self-reports of psychological well-being (Elliot et al., 2006; Horst et al., 2007; Ryan & Shim, 2006). However, we do not know if social achievement goals are an important aspect of social motivation for young adolescents and have implications for their social adjustment in school. In the following sections, we further describe social achievement goals, their likely relevance during early adolescence, and our hypotheses about their significance for social adjustment during middle school.

### A Social Development Goal During Early Adolescence

With its focus on improving social competence, a social development goal is expected to relate to adaptive beliefs and behaviors. The nature of peer relationships evolves markedly during early adolescence, and it is likely that a social development goal would be adaptive in successfully navigating the changing terrain of peer relationships during early adolescence. Just as thinking skills advance (Keating, 2004), conceptions of peer relationships, interpersonal understanding, and communication become more sophisticated during early adolescence (Eisenberg & Morris, 2004; Ford, 1982; Selman, 1980; Youniss & Smollar, 1985). Friendships are increasingly characterized by higher levels of empathy, self-disclosure, and responsiveness to others' thoughts and feelings (Newcomb & Bagwell, 1995; Savin-Williams & Berndt, 1990). In addition, many children make a transition to middle school that disrupts friendships and thrusts young adolescents into a much larger peer network (Eccles, 2004). A focus on developing peer relationships is likely to promote reflection on the evolving criteria of friendships during early adolescence and promote successful peer relationships.

As an approach goal (i.e., focusing on a positive outcome), a social development goal is likely to be undergirded by positive views of one's social competence and enhanced efficacy to achieve desired social outcomes (as found in other domains—see Elliot, 2005, for a review; and as found with adults in the social domain—see Ryan & Shim, 2006). Many social situations are likely to offer opportunities for developing social skills and friendships, and thus social development goals are likely to foster prosocial behavior and diminish aggressive behavior with peers. The quality of peer relationships is also likely to be enhanced. To achieve the goal of developing positive relationships, adolescents are more likely to attend to cues regarding compatibility with others, reflect on what is best for the relationship, and remember



information about their peers. These hypotheses are supported by research with adults that found a social development goal is related to positive perceptions of social relationships (Elliot et al., 2006; Horst et al., 2007; Ryan & Shim, 2006).<sup>1</sup> These hypotheses are also supported by research in the academic domain that has linked a mastery goal (analogous to a development goal in the social domain) to enhanced efficacy, preference for challenging and novel work, deep processing strategies, information retention, adaptive help seeking, and general engagement (see Elliot, 2005, for a review).

### Social Demonstration Goals During Early Adolescence

During adolescence, peer relationships also change regarding social comparison behavior and self-conscious emotions, which is relevant for social demonstration goals. In the context of school, students spend their time among peers and have much opportunity to observe and evaluate characteristics of their peers in a variety of situations. By early adolescence, students use social comparative information to evaluate their competence (Ruble, Boggianno, Feldman, & Loeb, 1980). Cognitive development also affords more adept perspective-taking abilities (Selman, 1980), which often results in young adolescents' unsettling realization that they are the focus of others' attention (termed the "imaginary audience"; Elkind, 1967). Compared to preadolescents and older adolescents, younger adolescents are more self-conscious (Simmons, Rosenberg, & Rosenberg, 1973). Simultaneously, during early adolescence students are increasingly focused on the emerging social phenomenon of popularity (high prestige and visibility among students at school; Eder, 1985; LaFontana & Cillessen, 2002; O'Brien & Bierman, 1988). Thus, given the increasing attention to comparisons and judgments of social attributes, it is likely that social demonstration goals are salient to many young adolescents.

### Social Demonstration-Approach Goal

A social demonstration-approach goal is conceptualized to be a more complex goal serving both approach tendencies (striving to achieve success, undergirded by high perceived social competence) and avoidance tendencies (concern with others' judgments and thus also undergirded by fear of failure; see Elliot & Church, 1997, for evidence in academic domain). Given the inherent approach nature of a social demonstration-approach goal, it will be positively associated with approaching and engaging in social situations with peers. Students who strive to achieve socially because they desire social status or recognition are likely to feel socially competent and interact with peers to achieve their goal (Ryan & Shim, 2006). However, the goal of demonstrating social superiority is likely to lead to aggression as well as prosocial behavior. Research has found that aggressive behavior is associated with achieving cool or popular status among peers (LaFontana & Cillessen, 2002; Parkhurst & Hopmeyer, 1998; Rodkin, Farmer, Pearl, & Van Acker, 2000; Rose, Swenson, & Waller, 2004).<sup>2</sup> Students who are perceived as popular by their peers also exhibit prosocial behavior, suggesting that a combination of aggressive and prosocial behaviors is effective in garnering social status (Hawley, 2003). These hypotheses are supported by research in the academic domain that has linked a performance-approach goal

(analogous to a social demonstration-approach goal) to engagement as well as disruptive behavior in the classroom (Kaplan, Gheen, & Midgley, 2002).

### Social Demonstration-Avoid Goal

With its focus on avoiding negative judgments, a social demonstration-avoid goal is expected to lead to maladaptive beliefs and behaviors. As an avoidance goal, it is expected to be undergirded by negative views of one's social competence and thus associated with diminished efficacy to achieve desired social outcomes (as found in other domains—see Elliot, 2005, for a review; and as found with adults in the social domain—see Ryan & Shim, 2006). Avoidance or withdrawal from social situations is likely to be preferred to engaging in social interactions, as it is safer and satisfies the goal of avoiding possible negative outcomes. Congruent with the focus on negative possibilities, a social demonstration-avoid goal is associated with social anxiety. The quality of peer relationships also suffers because a focus on negative outcomes and being overly self-conscious and afraid of failure is likely to undermine positive interactions with peers by increasing anxious or avoidant strategies and behaviors. Thus, a social demonstration-avoid goal is expected to hinder the formation of positive peer relationships. These hypotheses are supported by research with adults that links a social demonstration-avoid goal to negative perceptions of social relationships (Horst et al., 2007; Ryan & Shim, 2006), higher social worry (Ryan & Shim, 2006), increased fear of negative evaluation (Horst et al., 2007), and loneliness (Elliot et al., 2006). This is also in line with findings that negative approval-based self-appraisals are associated with increased social anxiety (Rudolph, Caldwell, & Conley, 2005). These hypotheses are also supported by research that has linked an academic performance-avoid goal (analogous to our social demonstration-avoid goal) to avoidant behavior and anxiety in academic situations (Middleton & Midgley, 1997; Ryan, Patrick, & Shim, 2005; Skaalvik, 1997).

### The Role of Perceived Social Competence

An additional aim of our research is to examine the role of perceived social competence in regard to social achievement goals and social adjustment. Perceived social competence refers to student beliefs about how good their skills and abilities are in the social domain (Harter, 1982). As we have described in our conceptualization, we expect a positive correlation between perceived social competence and both approach goals (social development

<sup>1</sup> Elliot et al. (2006) investigated "friendship-approach" and "friendship-avoid" goals that are conceptually similar to social development and social demonstration-avoid goals, but whose items focus specifically on friends and close relationships.

<sup>2</sup> Researchers have measured "popular" in different ways. One way is to ask children to list whom they like and whom they dislike in a classroom (referred to as sociometric status). Children with many like and few dislike nominations are popular and tend not to be aggressive. Another way is to ask children directly to identify who is popular or ask them to rate their own popularity (referred to as *perceived popularity*). There is a positive correlation between perceived popularity ratings and aggression (Cillessen & Rose, 2005).



and social demonstration-approach) and a negative association between perceived social competence and a social demonstration-avoid goal (see Elliot, 2005, for a review of other domains). Further, on the basis of research in social development, we expect that perceived social competence will be related to social adjustment (Harter, 1982; Ladd & Price, 1986; Robins & Pals, 2002; Wheeler & Ladd, 1982). Thus, perceived social competence is intimately related to the central constructs of this investigation. To fully understand the role of social achievement goals for social adjustment, we consider them in tandem with perceived social competence to investigate unique and possibly interactive effects. First, we examine social achievement goals and perceptions of social competence simultaneously as predictors of social adjustment in our regression analyses. We expect social achievement goals to predict social adjustment above and beyond perceptions of social competence. Such unique effects for achievement goals and perceived competence have been established in other domains (see Elliot, 2005, for a review) but have not been examined in the social domain. To support our contention that social achievement goals are important to social adjustment, it is important to show that they predict social adjustment above and beyond omnibus perceptions of social competence.

Second, we examine possible interactive effects between perceived social competence and social demonstration goals. Focusing on the judgments of others may have different implications depending on how confident someone is in his or her abilities. Although theory and evidence are mixed, this issue has been the source of much attention in the academic domain (Dweck & Leggett, 1988; Elliot & Church, 1997; Elliott & Dweck, 1988; Harackiewicz & Elliot, 1993; Smiley & Dweck, 1994). Dweck and Leggett's (1988) view is that perceived competence moderates the effects of demonstration goals (analogous to performance goals in the academic domain). Specifically, it is only when perceived competence is low that demonstration goals are associated with maladaptive outcomes (Dweck & Leggett, 1988). Dweck and colleagues' studies with children have found evidence for this view of perceived competence as a moderator (e.g., Elliott & Dweck, 1988; Smiley & Dweck, 1994). Elliot views perceived competence as related to achievement goals but does not view it as a moderator. Elliot and colleagues' research with college students consistently documents support for his view (e.g., Elliot & Church, 1997; Harackiewicz & Elliot, 1993). Thus, we examine whether perceived social competence moderates the relation of social demonstration goals to social adjustment.

### Interactive Effects Among Social Achievement Goals and by Gender

Although they are not the main focus of the present research, we explore two other types of interactive effects that have received attention in achievement goal theory and research. First, we examine whether gender moderates the effects of achievement goals on social adjustment (see Midgley, Kaplan, & Middleton, 2001). This is important because there is some research that suggests that achievement goals may have different implications for boys and girls, particularly performance goals (analogous to demonstration goals; e.g., Bouffard, Boisvet, Vezeau, & Larouche, 1995; Urdan, 1997). Much extant achievement goal research has not examined such gender by achievement goal interactions, so the extent of such

gender differences is not known. Second, we examine interactive effects among achievement goals. This is important because students may pursue multiple goals, and the effects of one goal may be different when pursued in conjunction with another goal (see Barron & Harackiewicz, 2001). The evidence for interactive effects among achievement goals is inconsistent, but again, many researchers have not examined interactions, which hinders definitive conclusions (Barron & Harackiewicz, 2001; Midgley et al., 2001). Of the three studies concerning social achievement goals in college students, one study examined potential interactions and few were found (Ryan & Shim, 2006). We make no hypotheses, but explore these types of interactions among social achievement goals in young adolescents to ensure we are not missing important patterns by focusing only on main effects. This is particularly important as we expand achievement goal theory into a new domain.

## Summary of the Present Research

In summary, we designed the present research to investigate the extension of an achievement goal framework to the social domain for young adolescents. In Study 1, we collected open-ended data on young adolescents' descriptions of social goals and social competence. In Study 2, we created new survey items to assess young adolescents' social achievement goals. We assessed the structure and nature of the new measure with factor analyses and examined the hypothesized relations of students' social achievement goals to their subsequent social adjustment (assessed with teacher and self reports).

### Study 1

In Study 1 we sought to establish that an achievement goal perspective is relevant and applicable to the social domain for young adolescent students. We used open-ended questions to elicit descriptions of social strivings from young adolescents. We used three different open-ended questions with subgroups within our sample, such that each child only responded to one question. The use of different questions bolsters confidence that results reflect an accurate portrayal of young adolescents' social strivings and are not an artifact produced by the use of a particular term or phrase in the question. The use of brief open-ended responses from a large and diverse sample of young adolescents (as opposed to lengthier interview responses from a much smaller number of students) increases confidence that results are actually representative and salient in young adolescents' social strivings. Showing that different orientations to social competence can be seen when students write a sentence or two will be convincing evidence that an achievement goal approach is relevant and potentially important to young adolescents' social adjustment.

### Method

#### Participants

The participants were sixth-grade students from two public middle schools. The schools serve nonmetropolitan, small urban communities and report that about half of their students are eligible



for free or reduced-price lunch.<sup>3</sup> At each school, we asked the principals to choose a few representative classrooms to participate in the research. Letters describing the project were sent home with all children in those classrooms, and 80% of the students returned a slip with their parents' written consent for them to participate. The sample ( $N = 153$ ) was 49% female and ethnically diverse (43% White, 35% Black, 8% Asian, 5% Hispanic, 3% Asian Indian, and 7% other).

### Procedure

Trained researchers distributed questionnaires to students in their classrooms. Teachers remained in the classrooms. The questions were read aloud, and then students were given time to write their response. Students were told there were no right or wrong answers and that the purpose of the questionnaire was to find out students' thoughts at school. Students were assured that the information would be kept confidential. In addition, students were told that answering the questions was voluntary; if at any point they wanted to stop, they could do so.

### Measures

Students responded to one of three different versions of the open-ended questions. Thirty-five students responded to Question 1 (Q1), which was, "What are your social goals for this year in school? Please elaborate why these goals are important to you." Sixty-one students responded to Question 2 (Q2), which had two parts. Part A was, "If you were a social success this year in sixth grade what would this look like? Please elaborate why you would feel like a success." Part B was, "If you were a social failure this year in sixth grade what would this look like? Please elaborate why you would feel like a failure." Fifty-seven students responded to Question 3 (Q3), which was, "Sometimes kids feel good about how they get along with other kids at school. Other times kids do not feel good about how they get along with other kids at school. Think about how you are with other kids at school. Now think about the rest of the school year. What are your goals for getting along with other kids at school? Please elaborate why these goals are important to you." For all three questions, students were told that by "social" we meant how you get along with other kids at school—friendships and social life in general at school. These three questions were developed from eight half-hour interviews with sixth-grade students (not part of the present sample).

### Results

All responses were coded independently by two coders. Students generally wrote one to two sentences below their question. We treated all information from each student as one response (thus, we had 153 distinct responses). Coding proceeded in two phases. First, we coded for evidence of development versus demonstration. Second, we coded for approach or avoid valence. Our conceptual framework did not concern a development-avoid goal because, in part, we assumed it would be less prevalent; by coding for this goal in the open-ended data, we could examine the assumption.

### Development Versus Demonstration

In Phase 1, we assigned each student response one of three codes: *development* (evidence of development only), *demonstration*

(evidence of demonstration only), or *neither* (response concerned social strivings but there was no evidence for either social achievement goal). Initially, we also included a category for a response that contained evidence of both development and demonstration goals, but there were few students in this category—5% overall, 2 (6%) for Q1, 2 (3%) for Q2, and 4 (7%) for Q3. As the aim of the study was to provide evidence that social achievement goals were relevant to social strivings in general (not to compare configurations in individuals), this category was dropped. Instead, if a student response contained reference to both social development and demonstration goals, it was double coded (about 5% of students had responses that were double coded). Fifteen students wrote responses that contained no reference to the social domain and thus were considered missing data and removed from the coding process (15 overall; 10, 0, and 5, for Q1, Q2, and Q3, respectively; sample numbers reflect final numbers of students with valid data). Reliability between coders was 90% ( $\kappa = .85$ ,  $p < .001$ ; Cohen, 1960). Disagreements were resolved through discussion.

A response was coded as social development if it concerned a focus on developing social competence. To be coded as evidence of a social development goal, comments needed to go beyond achieving a certain outcome and make explicit reference to improvement. Thus, comments such as "develop better friendships," "be more friendly," and "be less selfish" were considered evidence for a social development goal, whereas comments such as "make friends," "be nice," or "have fun" were not (and thus were coded as "neither achievement goal"). A response was coded as evidence for a social demonstration goal if it concerned a focus on others' judgments of social competence and how one looked in front of others. Thus, comments such as "I would be seen as having lots of friends" and "Many people would like me" were considered evidence of a social demonstration goal because they explicitly refer to others' impressions, whereas comments such as "I would have lots of friends" were not. Further, comments such as "I would be popular," "I would be cool," "I would have popular friends," or "I would have cool friends" were coded as a social demonstration goal because use of the terms popular or cool connotes a focus on social recognition from others and a concern with high social status or a social standing that is better than others (i.e., the nonpopular or uncool people). Examples of social development and demonstration goals for each of the three questions are shown in Table 1.

Table 2 shows the percentages for the different codes. Across all three questions, 70 of the students (46%) made reference to social development goals, 54 students (35%) made reference to social demonstration goals, and 40 students (26%) gave responses that did not contain strong evidence for either social development or social demonstration goals. There were some differences in the results by question. Q2 garnered more demonstration than development responses, whereas Q1 and Q3 garnered more development than demonstration responses. However, both types of ori-

<sup>3</sup> These schools do not separate students by gifted or special education status. In both Studies 1 and 2, a few special education students were excluded because their teacher judged the research inappropriate for their skill level.

Table 1  
*Examples of Social Development and Social Demonstration Achievement Goals From Open-Ended Data*

Open-ended questions	Social development goal	Social demonstration goal
What are your social goals for the upcoming year? Please elaborate why these goals are important to you.	I want to grow my friendships with my friends. Be kinder than I am now. Be a better friend. Know my friends better.	Be popular so when I walk in the hallway everybody knows my name. Not be picked on. I don't want to be made fun of
If you were socially successful next year what would this look like? Please elaborate why you would feel like a success.	I would be nicer to others. I made friends and learned more about people. If I made new friends... learn to open up when I need to and I will know when to stop.	Everyone will talk to you. You would feel like the center of attention all the time. A lot of people would know me... People would recognize me when I walked through the hallway. I would be cool... feel like a success because I was popular and a cool kid, not one of those nerds. I had a new girl friend and made tons of new friends and sat at the popular table.
If you were a social failure next year what would this look like? Please elaborate why you would feel like a failure.	I would feel lonely and without any friends... I would feel like I didn't do my best. I would not care about anyone but myself and not let people get to know me or get to know other people. I would have no one I could trust and I would be letting myself down.	I would be a social outcast, little or nobody would like me. I'd feel like a loser, not popular. I would have no friends, be at the bottom of the social pyramid. People would laugh at me.
Sometimes kids feel good about how they get along with other kids at school. Other times kids do NOT feel good about how they get along with other kids at school. Think about how you are with other kids at school. Now think about the rest of the school year. What are your goals for getting along with other kids at school?	Be funnier. Be more open so I can learn about others. Become a better friend to others. Act in better/kinder ways for my friends. Learn more about all my friends. Be less selfish. Try to get along with Warren better. We've been friends for 3 years but he can still go into flurries of rage sometimes.	Get Jessie Rubins to like me... she's rather popular. I want to have a good reputation. Get respect. Not have no one like me. Become more popular. Not be picked on by others... be recognized by others.

*Note.* Students were told that by "social" we meant social skills, social relationships, and social life in general.

entation to social competence could be seen in responses to all three questions.

### *Approach Versus Avoid*

In phase two of the coding, each coded response for Q1 and Q3 was then coded for approach-avoid valence. Q2 was not coded for approach-avoid valence because we had explicitly asked students to describe success (approach goal) and failure (avoid goal). We assigned each student response to Q1 and Q3 one of four codes: approach (striving to achieve a positive outcome), avoid (striving

to avoid a negative outcome), both (containing both approach and avoid goal statements), or neutral (desiring to stay the same and vague about whether this would be continuing to accomplish a positive outcome or continuing to avoid a negative outcome). For example, "be more generous" would be coded as approach, "be less selfish" would be coded as avoid, and "stay the same" would be coded as neutral. For the 55 responses that had been coded as having evidence of a social development goal in Q1 and Q3, 65% (38) were approach, 15% (8) were avoid, 13% (7) contained both, and 7% (4) were neutral. For the 21 responses that were coded as having evidence of a social demonstration goal in Q1 and Q3, 48%

Table 2  
*Open-Ended Responses Coded as Evidence of Development, Demonstration, or Neither Social Achievement Goal*

Goal	Question 1 (n = 35)		Question 2 (n = 61)		Question 3 (n = 57)		Total (N = 153)	
	Response	%	Response	%	Response	%	Response	%
Development	20	57	15	25	35	61	70	46
Demonstration	6	17	33	54	15	26	54	35
Neither	11	31	17	28	12	21	40	26

*Note.* Percentages do not add to 100% because some students' responses made reference to both development and demonstration goals and thus were double coded.



(10) were approach, 24% (5) were avoid, 29% (6) contained both, and 0% were neutral. Across questions, reliability between coders was 100% (number of agreements/number of possible agreements).

### Discussion

Results from Study 1 indicate that different orientations toward developing or demonstrating social competence can be seen in young adolescents' descriptions of their social goals and evaluations of social competence. The majority of students had comments that could be coded as evidence of a social development or social demonstration goal orientation. Different questions yielded different percentages of development versus demonstration goals, but evidence of both was found in each question; looking at responses across the questions, there was clear evidence for both types of social achievement goals. The fact that different orientations to social competence could be seen when students wrote a sentence or two is convincing that this is relevant and potentially important. Study 1 also documented both approach and avoidance goals in students' social strivings. It is interesting that there was a higher percentage of approach than avoid goals, and this was particularly true for the development goals. This is consistent with theories that mastery-avoid goals are less prevalent (Elliot, 2005). This result supports our decision not to include a development-avoid goal in this initial investigation of social achievement goals for young adolescents, but given that there were examples of development-avoid goals, it may be important for future research to attend to them. Overall, Study 1 established that an achievement goal framework is relevant to young adolescents' social motivation.

### Study 2

In Study 2, informed by Study 1, we created new survey items to assess individual differences in young adolescents' social achievement goals. Study 2 was guided by several key aims. First, we examined whether the three social achievement goals (development, demonstration-approach, demonstration-avoid) are distinct and reliable constructs. Given the new nature of the measure, we used exploratory factor analysis. Second, we examined the correlations among the social achievement goals to better understand the nature of social achievement goals in young adolescents. We expected a moderate correlation between goals that shared a focus on demonstration (i.e., demonstration-approach and demonstration-avoid goals; Horst et al., 2007; Ryan & Shim, 2006). Third, we examined the correlations between perceived social competence and social achievement goals. We expected perceived social competence to be positively related to both approach goals (development and demonstration-approach goals) and negatively related to demonstration-avoid goals (Ryan & Shim, 2006).

Next, we examined the relation of social achievement goals to social adjustment. Consistent with our conceptualization that goals precede social adjustment, we used a prospective longitudinal design to examine whether social achievement goals foreshadow subsequent social adjustment (3-month time span). We expected that different social achievement goals would set in motion different ways of engaging in the social world and have different

implications for social adjustment. In line with the focus on developing social skills and relationships, a social development goal would be positively associated with subsequent prosocial behavior and negatively associated with subsequent aggressive behavior. In line with the focus on demonstrating social desirability relative to others, a social demonstration-approach goal would be positively associated with both prosocial and aggressive behavior. In line with the focus on avoiding negative judgments from others, a social demonstration-avoid goal would be positively associated with subsequent anxious solitary behavior and social worry. Regarding friendship features and perceived popularity, a social development goal would be the strongest predictor of positive features in close friendships, whereas a social demonstration-approach goal would be the strongest predictor for perceived popularity. As approach goals, they both might have positive associations with positive outcomes, although it is expected that the relations representing closer alignment of goal content and outcome (i.e., development-quality friendship features and demonstration-approach-popularity) would be stronger. A social demonstration-avoid goal is expected to have negative associations with both friendship features and popularity, although the latter is likely to be stronger in magnitude, as concern about negative judgments may be magnified in public rather than private and more predictable peer settings.

To assess social adjustment, we used both self-reports and teacher reports. Student perceptions are an important aspect of their adjustment (Achenbach, McConaughtly, & Howell, 1987), and much of the research on academic achievement goals of young adolescents has relied on student reports of adjustment (with the exception of report card grades; e.g., Kaplan et al., 2002; Middleton & Midgley, 1997; Skaalvik, 1997). However, given the limitations associated with self-reports (e.g., common method variance explanation for effects), we also asked teachers to assess the aspects of social adjustment that would be visible to them (i.e., prosocial, aggressive, and anxious solitary behavior). Thus, a strength of the present research is the examination of social achievement goals and social adjustment beyond just self-report measures to establish construct validity.

### Method

#### Participants

The participants were sixth-grade students from the same two public middle schools ( $n = 217$  at Wave 1 in December and  $n = 196$  at Wave 2 in March and April) that were used in Experiment 1. Of the 217 students who had Wave 1 data, 181 had Wave 2 self-report data, and 208 had Wave 2 teacher-report data. Thus, the sample for analyses involving self-report social adjustment outcomes was 181, and the sample for analyses involving teacher-report social adjustment outcomes was 208. About two thirds of the sample had previously participated in Study 1. Letters describing the project were sent home, and 84% of the students returned a slip with their parents' written consent for them to participate (resulting in the sample described above). Students reported their gender and ethnicity. The sample was 47% female and ethnically diverse (39% Black, 35% White, 11% Asian, 5% Asian Indian, 4% Hispanic, and 6% other).

## Procedure

Surveys were administered to students in their classrooms by trained researchers. Teachers remained in the classrooms. Social achievement goals and perceived social competence were measured at Wave 1 (December), and social adjustment indices were measured about 3 months later at Wave 2 (March–April). Instructions and items were read aloud while students read along and responded. Students were told that the survey was not a test, that there were no right or wrong answers, and that the purpose of the survey was to find out students' beliefs and behaviors at school. Students were assured that the information would be kept confidential. In addition, students were told that filling out the survey was voluntary; if at any point they wanted to stop, they could do so.

Teachers were asked to complete a brief survey about students' behavior in school at Wave 2. Teachers were told that the purpose of the survey was to better understand students' adjustment to middle school; their participation was voluntary and their responses confidential. In middle schools, students have several different teachers throughout the school day. Thus, we had two different teachers (of main academic subjects) provide independent assessments for a more complete assessment of student behavior in middle school.

## Measures

**Social achievement goals.** We developed new items to measure the three social achievement goals during early adolescence. Social development goal items focus on developing social competence. Social demonstration-approach goal items focus on demonstrating social desirability and gaining positive judgments from others. Social demonstration-avoid goal items focus on demonstrating that one is not socially undesirable and avoiding negative

judgments from others. We reviewed the social goal items Ryan and Shim (2006) used with college students, as well as results of Study 1, to generate items that would be appropriate for young adolescents. We did some pilot testing of a larger pool of potential social goal items for young adolescents and ultimately settled on the 18 goal items reported in Table 3. Conceptually, the new items were parallel to Ryan and Shim, but the wording needed to be different to be appropriate for young adolescents. Of our 18 items, 2 items were identical, 3 items were slightly adapted, and 13 items were new compared to the scale used by Ryan and Shim for college students. All items were rated on a scale that ranged from 1 (*not at all true of me*) to 5 (*very true of me*). Factor and reliability analysis is presented in the Results section.

**Perceived social competence.** Eccles's measure of perceived competence for the social domain was used (Lord, Eccles, & McCarthy, 1994). There were four items, which were rated on a scale that ranged from 1 (*not at all good*) to 5 (*very good*). Sample items are "Compared to most sixth-grade students, how would you rate your social skills?" and "How good are you at making friends?" The measure was reliable in our sample ( $\alpha = .82$ ).

**Prosocial behavior.** Prosocial behavior was assessed with a measure adapted from Cassidy and Asher (1992) and Crick (1996), and consisted of five items: "friendly," "helpful," "cooperative," "kind," and "considerate." Teachers and students rated the items on a scale that ranged from 1 (*never*) to 5 (*always*). The correlation between the two teacher ratings for each student was  $r = .51, p < .05$ . Students' score on the teacher-reported measure was the average of the two teacher reports. The measure was reliable in our sample ( $\alpha = .91$  and  $.84$ , for teacher-reported scale and student-reported scale, respectively).

**Overt aggressive behavior.** Overt aggressive behavior was assessed with the Aggression subscale of the Interpersonal Com-

Table 3  
Factor Loadings for Social Achievement Goals

Item	Social development	Social demonstration-approach	Social demonstration-avoid
I like it when I learn better ways to get along with friends.	.87		
I feel successful when I learn something new about how to get along with other kids.	.83		
I try to figure out what makes a good friend.	.73		
One of my goals is that my friendships become even better over time.	.71		
It is important to me to learn more about other kids and what they are like.	.64		
In general, I try to develop my social skills.	.62		
It is important to me that other kids think I am popular.		.90	
It is important to me to have "cool" friends.		.85	
I want to be friends with the "popular" kids.		.75	
It is important to me to be seen as having a lot of friends.		.75	
I try to do things that make me look good to other kids.		.62	
My goal is to show other kids how much everyone likes me.		.55	
I try not to do anything that might make other kids tease me.			.74
It is important to me that I don't embarrass myself around my friends.			.72
I try to avoid doing things that make me look foolish to other kids.			.66
When I am around other kids, I don't want to be made fun of.			.59
When I am around other kids, I mostly just try not to goof up.			.53
One of my main goals is to make sure other kids don't say anything bad about me.			.47
Eigenvalue	2.42	7.32	1.57
% of variance explained	11.17	38.28	6.23
$\alpha$	.87	.89	.84

Note. Factor loadings above .30 are shown.



petence Scale (Cairns, Leung, Gest, & Cairns, 1995) and consisted of three items: "fights with others," "argues with others," and "gets in trouble." Teachers and students rated the items on a scale that ranged from 1 (*never*) to 5 (*always*). The correlation between the two teacher ratings for each student was  $r = .71, p < .05$ . Students' score on the teacher-reported measure was the average of the two teacher reports. The measure was reliable in our sample ( $\alpha = .88$  and  $.73$ , for teacher-reported scale and student-reported scale, respectively).

*Anxious solitary behavior.* Anxious solitary behavior was adapted from Gazelle's measure (Gazelle & Ladd, 2003; Gazelle & Rudolph, 2004) and consisted of six items: "worries," "anxious," "self-conscious," "shy and timid," "withdrawn from peers," and "prefers to be alone." Teachers and students rated the items on a scale that ranged from 1 (*never*) to 5 (*always*). The correlation between the two teacher ratings for each student was  $r = .69, p < .05$ . Students' score on the teacher-reported measure was the average of the two teacher reports. The measure was reliable in our sample ( $\alpha = .81$  and  $.72$  for teacher-reported scale and student-reported scale, respectively).

*Social worry.* We assessed students' worry regarding their social behavior and relationships with a measure used by Ryan and Shim (2006). Students were directed to write a number from 1 (*not at all true of me*) to 5 (*very true of me*) on the line next to various statements about academic and social worries. The four statements regarding social worries (e.g., "I worry about what my friends think of me" and "Many social situations make me worry") were averaged together such that a high score indicated higher levels of social worry. The measure was reliable in our sample ( $\alpha = .83$ ).

*Positive friendship quality.* We used Rose's (2002) adapted version of Parker & Asher's (1993) Friendship Quality Questionnaire to measure students' perceptions of their positive relationship qualities with their best friend at school. To facilitate a focus on one best friendship, we instructed students to write down the name of their very best friend at school and think of that best friend as they completed items about support (e.g., "Gives advice with figuring things out," 3 items), validation (e.g., "Makes me feel good about my ideas," 2 items), intimacy (e.g., "We can talk about whatever happens to us," 3 items), and conflict resolution (e.g., "We can talk about how to get over being mad at each other," 2 items). All items were rated on a scale that ranged from 1 (*never*) to 5 (*always*). These 10 items formed one factor in a factor analysis and thus were averaged to form one scale. The measure was reliable in our sample ( $\alpha = .89$ ).

*Popularity.* We measured perceived popularity with the Popularity subscale of the Interpersonal Competence Scale (Cairns et al., 1995) consisting of three items ("popular with boys," "popular with girls," and "has lots of friends"). All items were rated on a scale that ranged from 1 (*never*) to 5 (*always*). The measure was reliable in our sample ( $\alpha = .78$ ).

## Results

### Exploratory Factor Analyses

Since these were new items that had never been tested in this age group before, we conducted an exploratory factor analysis. A principal axis factor analysis with oblimin rotation was conducted for the entire sample ( $N = 217$ ) on the 18 social goal items at

Wave 1. Oblimin rotation was appropriate because we assumed the factors would be correlated. Factors whose eigenvalues were greater than 1 were extracted. The analysis yielded three factors that accounted for 55.69% of the total variance (see Table 3). The three factors corresponded to the three hypothesized social goals: social development, social demonstration-approach, and social demonstration-avoid goals. All factor loadings were above .47 on their primary factor. No items loaded on another factor at greater than .30.

### Reliability Analyses

Reliability analyses indicated that the three social achievement goal scales had good internal consistency (see Table 3). The six items in each of the three scales were averaged to create three social achievement goal scales with a range of 1 through 5.

### Correlations Between Social Achievement Goals and Perceived Social Competence

As expected, there was a correlation between the two demonstration goals ( $r = .57, p < .001$ ; see Table 4). As expected, there was a correlation between the two approach goals ( $r = .36, p < .01$ ). Unexpectedly, there was also a correlation between development and demonstration-avoid goals ( $r = .44, p < .001$ ). As expected, perceived social competence was related to both a development goal ( $r = .40, p < .01$ ) and a demonstration-approach goal ( $r = .28, p < .01$ ). In contrast to the hypothesized negative relation between perceived social competence and a demonstration-avoid goal, a positive relation was documented ( $r = .16, p < .05$ ).

### Regression Analyses: Examining Social Achievement Goals as Predictors of Subsequent Social Adjustment (Self-Report Measures)

*Overview of analyses.* Separate multiple regression analyses were conducted for each of the six self-report dependent variables: prosocial behavior, aggressive behavior, solitary anxious behavior, social worry, best-friend quality, and popularity (see Table 5). Social achievement goals and perceived social competence (measured at Wave 1), as well as gender, were entered as predictor variables. Preliminary analyses tested interactions between social demonstration goals and perceived social competence as well as interactions among social goals (i.e., Development  $\times$  Demonstration-Approach, Development  $\times$  Demonstration-Avoid, and Demonstration-Avoid  $\times$  Demonstration-Approach). Additionally, we examined whether gender moderated any of the effects. We used procedures outlined by Aiken and West (1991) to test and interpret interactions. Main effect terms were standardized before computing interaction terms to avoid multicollinearity and aid interpretation of beta coefficients. Significance of interactions was determined if the  $R^2$  increased by a significant amount and if the beta coefficient for the interaction term was significant as well. Only significant interaction terms (by both criteria) were retained in the final models (thus the final models varied for different dependent variables). To interpret the significant interactions, we calculated the predicted values using unstandardized regression coefficients and conducted simple slope tests. Graphs were created

Table 4

*Correlations Between Social Achievement Goals and Perceived Social Competence (Wave 1) and Social Adjustment (Wave 2)*

Measure	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Social development (SR)													
2. Social dem-ap. (SR)	.36												
3. Social dem-av. (SR)	.44	.57											
4. Perc'd social comp. (SR)	.40	.28	.16										
5. Prosocial (SR)	.30	-.10	.20	.37									
6. Aggressive (SR)	-.17	.14	-.10	-.08	-.50								
7. Solitary anxious (SR)	.06	.01	.16	-.23	.03	-.03							
8. Social worry (SR)	.18	.20	.31	-.02	-.06	.04	.37						
9. Best-friend quality (SR)	.32	.07	.09	.41	.37	-.10	.05	.04					
10. Perceived popularity (SR)	.34	.35	.09	.63	.16	.09	-.22	.04	.34				
11. Prosocial (TR)	.01	-.18	.02	.11	.31	-.32	.05	.06	.08	-.12			
12. Aggressive (TR)	.05	.20	-.08	.05	-.33	.41	-.07	-.03	-.06	.28	-.68		
13. Solitary anxious (TR)	-.03	-.14	.15	-.15	.10	-.07	.20	.02	-.07	-.24	.12	-.03	
<i>M</i>	3.78	2.94	3.35	4.15	3.80	2.28	2.43	2.32	3.71	3.75	3.90	1.89	2.13
<i>SD</i>	0.92	1.10	0.99	0.78	0.70	0.82	0.72	0.94	0.90	1.01	0.67	0.88	0.77

Note. Correlations above .14 are significant at .05 level; above .18 are significant at .01 level. SR = self-report measures; dem-ap. = demonstration-approach; dem-av = demonstration-avoid; perc'd social comp. = perceived social competence; TR = teacher-report measures.

using the predicted values of outcome variables for the students with goal scores one standard deviation above and below the mean.

**Prosocial behavior.** The regression of prosocial behavior on the model yielded significant effects for the final model,  $F(5, 169) = 12.48, p < .001, R^2 = .27$ . There were significant relationships for development and demonstration-approach goals but not for a demonstration-avoid goal. A development goal was positively related to prosocial behavior ( $\beta = .30, p < .001$ ), whereas a demonstration-approach goal was negatively related to prosocial behavior ( $\beta = -.40, p < .001$ ). Perceived social competence was also a significant predictor of prosocial behavior ( $\beta = .36, p < .001$ ).

**Aggressive behavior.** The regression of aggressive behavior on the model yielded significant effects for the final model,  $F(10, 164) = 3.93, p < .001, R^2 = .17$ . There were main effects for each of the three goals that were qualified by significant interaction terms. A significant interaction between a social demonstration-approach and a development goal ( $\beta = -.22, p < .05$ ) indicated that the positive relation between a social demonstration-approach goal and aggressive behavior was not significant when a social development goal was high (see Figure 1). There was also a three-way interaction between a social demonstration-avoid goal, perceived social competence, and gender ( $\beta = .27, p < .01$ ). For girls with high levels of perceived social competence, a social

Table 5

*Standardized Regression Coefficients for Predicting Self-Reports of Social Adjustment at Wave 2 From Social Achievement Goals and Perceived Social Competence at Wave 1*

Variable	Prosocial behavior	Aggressive behavior	Anxious solitary behavior	Social worry	Best-friend quality	Perceived popularity
DEV	.30***	-.20*	.14	-.06	.17*	.07
DAP	-.40***	.27**	-.06	.08	-.05	.28***
DAV	.14	-.30**	.20*	.22*	-.01	-.22**
PSC	.36***	-.10	-.32***	-.12	.33***	.55***
Gender <sup>a</sup>	.03	-.07	.05	.14	.24**	.05
DEV × DAP		-.22*				
DEV × DAV						-.13*
DAP × DAV				-.27***		
DAV × PSC × Gender		.27**				
DAV × PSC		.06				
DAV × Gender		-.07				
PSC × Gender		.07				

Note. As suggested by Aiken and West (1991), only significant interaction terms were retained in final models. Lower order two-way interaction terms were included to properly examine the three-way interaction term. Thus, final models varied for each outcome. The standardized regression coefficients shown are from the final model with all main effects and interaction terms in the model. When the interaction terms were added to the models, they explained 7%, 7%, and 2% additional variance (all significant at  $p < .05$ ) for aggressive behavior, social worry, and perceived popularity, respectively. DEV = social development goal; DAP = social demonstration-approach goal; DAV = social demonstration-avoid goal; PSC = perceived social competence.

<sup>a</sup> Gender is coded 1 = female and 0 = male.

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .



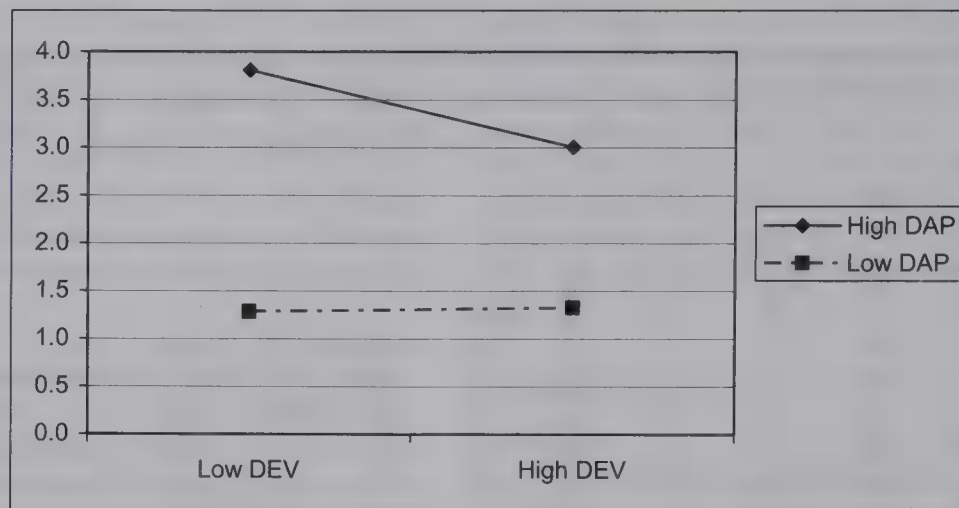


Figure 1. Social Demonstration-Approach (DAP)  $\times$  Development (DEV) goal interaction on aggressive behavior.

demonstration-avoid goal was positively related to subsequent aggressive behavior. For girls with low levels of perceived social competence and boys of all levels of perceived social competence, a social demonstration-avoid goal was negatively related to subsequent aggressive behavior (see Figure 2).

**Anxious solitary behavior.** The regression of anxious solitary behavior on the model yielded significant effects for the final model,  $F(5, 157) = 4.15, p < .01, R^2 = .12$ . There was a significant relationship for a demonstration-avoid goal, but not for

development and demonstration-approach goals. A demonstration-avoid goal was positively related to anxious solitary behavior ( $\beta = .20, p < .05$ ). Perceived social competence was also a significant negative predictor of anxious solitary behavior ( $\beta = -.32, p < .001$ ).

**Social worry.** The regression of social worry on the model yielded significant effects for the final model,  $F(6, 167) = 6.69, p < .001, R^2 = .19$ . There was a main effect for a demonstration-avoid goal that was qualified by a significant interaction term ( $\beta =$

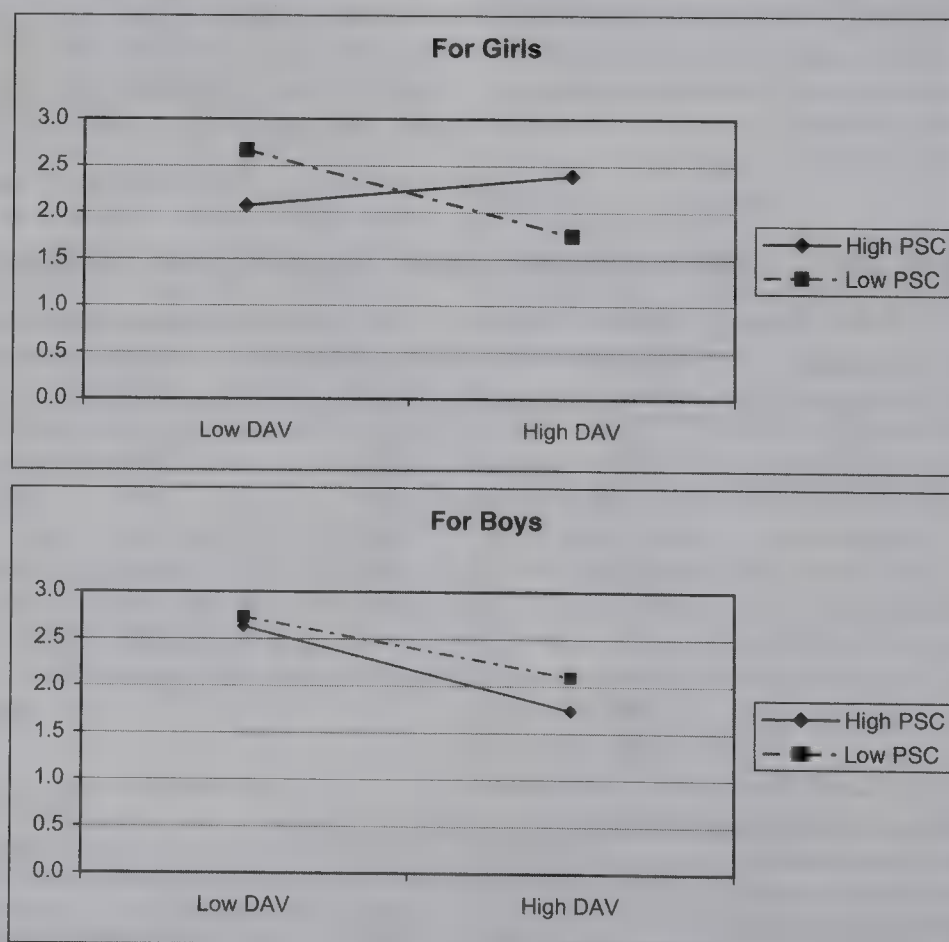


Figure 2. Social Demonstration-Avoid (DAV)  $\times$  Perceived Social Competence  $\times$  Gender interaction on aggressive behavior. PSC = perceived social competence.

-.27,  $p < .001$ ). A significant Demonstration-Avoid Goal  $\times$  Demonstration-Approach Goal interaction indicated that the positive relation between a social demonstration-avoid goal and social worry was only found when a social demonstration-approach goal was low; there was no relation when a social demonstration-approach goal was high (see Figure 3). However, as can also be seen in Figure 3, a social demonstration-approach goal alone can also result in high social worry. The only significant difference between groups was that students low in both demonstration goals were lower than the other groups. Thus, students with a high level of one or both of the demonstration goals had similar levels of worry, which was higher compared to students with low levels of both types of demonstration goals.

**Best-friend quality.** The regression of best-friend quality on the model yielded significant effects for the final model,  $F(5, 166) = 11.13$ ,  $p < .001$ ,  $R^2 = .25$ . There was a significant relationship for a development goal, but not for demonstration-approach and demonstration-avoid goals. A development goal was positively related to best-friend quality ( $\beta = .17$ ,  $p < .05$ ). Perceived social competence was also a significant predictor of best-friend quality ( $\beta = .33$ ,  $p < .001$ ), as was gender ( $\beta = .24$ ,  $p < .01$ ).

**Perceived popularity.** The regression of perceived popularity on the model yielded significant effects for the final model,  $F(6, 168) = 25.20$ ,  $p < .001$ ,  $R^2 = .47$ . A demonstration-approach goal was positively related to perceived popularity ( $\beta = .28$ ,  $p < .001$ ). There was a significant interaction between development and demonstration-avoid goals ( $\beta = -.13$ ,  $p < .05$ ). A social development goal was related to subsequent perceptions of popularity. However, this hypothesized positive relation was only found when a social demonstration-avoid goal was low; there was no relation when a social demonstration-avoid goal was high (see Figure 4). Perceived social competence was also a significant predictor of perceived popularity ( $\beta = .55$ ,  $p < .001$ ).

#### Regression Analyses: Examining Social Achievement Goals as Predictors of Subsequent Social Adjustment (Teacher-Report Measures)

**Overview of analyses.** Separate multiple regression analyses were conducted for each of the three teacher-report dependent variables with the same model-building procedures as described for the self-reported adjustment indices. Most of the main effects

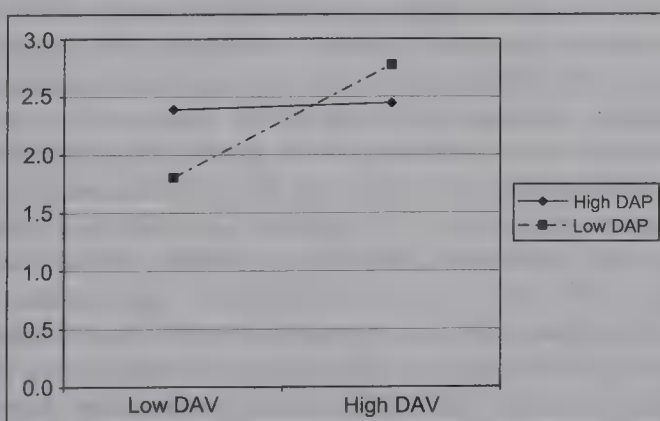


Figure 3. Social Demonstration-Avoid (DAV)  $\times$  Demonstration-Approach (DAP) interaction on social worry.

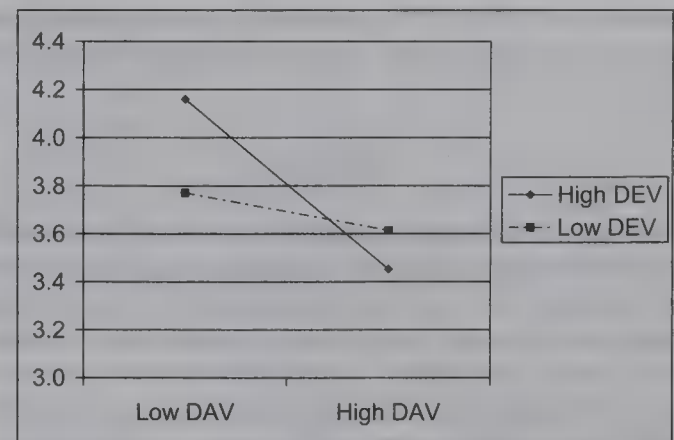


Figure 4. Social Development (DEV)  $\times$  Demonstration-Avoid (DAV) goal interaction on perceived popularity.

found for self reports of social adjustment were replicated with teacher reports of social adjustment (see Table 6).

**Prosocial behavior.** The regression of teacher-reported prosocial behavior on the model yielded significant effects for the final model,  $F(5, 194) = 2.86$ ,  $p < .05$ ,  $R^2 = .07$ . There were significant relationships for a demonstration-approach goal but not for development or demonstration-avoid goals. A demonstration-approach goal was negatively related to prosocial behavior ( $\beta = -.26$ ,  $p < .01$ ). Perceived social competence ( $\beta = .15$ ,  $p < .05$ ) was also a significant predictor of prosocial behavior.

**Aggressive behavior.** The regression of teacher-reported aggressive behavior on the model yielded significant effects for the final model,  $F(5, 194) = 3.26$ ,  $p < .01$ ,  $R^2 = .08$ . There were significant relationships for demonstration-approach and demonstration-avoid goals, but not for a development goal. A demonstration-approach goal was positively related to aggressive behavior ( $\beta = .31$ ,  $p < .01$ ), whereas a demonstration-avoid goal was negatively related to aggressive behavior ( $\beta = -.27$ ,  $p < .01$ ).

**Anxious solitary behavior.** The regression of teacher-reported anxious solitary behavior on the model yielded significant effects for the final model,  $F(5, 194) = 2.72$ ,  $p < .01$ ,  $R^2 = .07$ . There were significant relationships for demonstration-approach and demonstration-avoid goals, but not for a development goal. A demonstration-avoid goal was positively related to anxious solitary

Table 6

Standardized Regression Coefficients for Predicting Teacher Reports of Social Adjustment at Wave 2 From Social Achievement Goals and Perceived Social Competence at Wave 1

Variable	Prosocial behavior	Aggressive behavior	Anxious solitary behavior
DEV	-.03	.06	.01
DAP	-.26**	.31**	-.24**
DAV	.13	-.27**	.21*
PSC	.15*	-.02	-.13
Gender <sup>a</sup>	.10	-.07	-.03

Note. DEV = social development goal; DAP = social demonstration-approach goal; DAV = social demonstration-avoid goal; PSC = perceived social competence.

<sup>a</sup> Gender is coded 1 = female and 0 = male.

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .



behavior ( $\beta = .21, p < .01$ ), whereas a demonstration-approach goal was negatively related to anxious solitary behavior ( $\beta = -.24, p < .05$ ).

### Discussion

The results from Study 2 supported the hypothesized model of social achievement goals. A face-valid measure of social achievement goals was developed and administered to young adolescent students. As expected, factor analyses indicated a three-factor model (social development, social demonstration-approach, and social development-avoid goals). Reliability analyses indicated that the social achievement goal measures have good internal consistency. Social achievement goals were important in explaining subsequent social adjustment. Effects for social achievement goals were independent of perceived social competence and gender. As expected, a distinct pattern of effects was found for each social achievement goal, consistent with the premise that different achievement goals set in motion different cognitive, affective, and behavioral processes. For several aspects of social adjustment, there were significant interactions among goals, indicating that social achievement goals must be considered jointly for a full understanding of their implications for social adjustment.

Results with self-report measures of social adjustment were consistent with the hypothesis that a social development goal would be associated with increased prosocial behavior, decreased aggressive behavior, and increased perceptions of positive qualities in close friendships. A focus on developing social competence seems to be a positive orientation toward the social world that sets in motion adaptive beliefs and behaviors. A social development goal was also related to subsequent perceived popularity, but only when a social demonstration-avoid goal was low, indicating that when a student is focused on both goals, the drawbacks of a social demonstration-avoid goal neutralize the benefits of a social development goal. It is interesting that a social development goal was not associated with teacher ratings of social behavior, as expected. Teachers and students attend to and see different things in the classroom, so this finding may reflect that prosocial behavior associated with a social development goal is more prevalent in peer interactions in situations that are not salient to teachers. Future work could examine peer ratings of prosocial behavior to determine if development goals manifest themselves in certain patterns of behavior that are observed by peers.

A social demonstration-approach goal was positively associated with aggressive behavior and negatively associated with prosocial behavior. This pattern was found for both self and teacher reports of behavior. As expected, a social demonstration-approach goal was positively related to perceived popularity. Collectively, these results suggest that the pursuit of positive judgments by peers as cool or popular may be associated with unprincipled behavior. Interestingly, for self reports, a social development goal ameliorated the positive relation between a social demonstration-approach goal and aggressive behavior. This suggests that an additional benefit of a social development goal is that it can minimize the aggressive behavior that is associated with a social demonstration-approach goal.

The unexpected negative relation between a social demonstration-approach goal and prosocial behavior in the regression models is troubling. This pattern suggests that "nice" behavior

is incompatible with the climb to the top of the social pyramid in middle school. According to both self and teacher reports of behavior, when students are focused on demonstrating social desirability, they are less likely to act in helpful, cooperative, and kind ways toward their peers. Previous research has found that high-status children display both aggressive and prosocial behavior (Hawley, 2003; LaFontana & Cillessen, 2002; Rose et al., 2004). The different results found in our study may reflect that we are examining behaviors associated with the goal for high social status, not behavioral correlates of social status already achieved. A contribution of the present research to the emerging literature on aggressive high-status youth is the focus on antecedent psychological variables of social status. An achievement goal perspective is informative about the developmental precursors of different social behaviors in achieving status in school.

A social demonstration-avoid goal was positively associated with subsequent anxious solitary behavior. This was true for both self and teacher reports of anxious solitary behavior, which provides convincing evidence that a focus on avoiding negative judgments from peers is associated with social behaviors that undermine social adjustment in middle school. Further, endorsement of a social demonstration-avoid goal was also negatively associated with subsequent self-ratings of popularity. A social demonstration-avoid goal had a null relation with positive features in close friendships, indicating that in familiar, private friend relations, an avoid orientation did not undermine the development of qualities such as intimacy and support.

An interaction between a social demonstration-avoid and a social demonstration-approach goal indicated that a focus on either goal was associated with increased social worry. It is interesting that there seems to be a threshold effect whereby focusing on both types of demonstration goals does not relate to higher levels of social worry compared to students who focused on only demonstration-approach or demonstration-avoid goals. Thus, when students are focused on others' opinions of them, that focus is associated with social worry, regardless of whether they are oriented toward garnering positive judgments or avoiding negative judgments, or both.

There was another interaction involving a social demonstration-avoid goal, perceived social competence, and gender. A social demonstration-avoid goal was related to aggressive behavior, although the relation was moderated by the level of perceived social competence and gender. For socially confident girls, a social demonstration-avoid goal was positively associated with aggressive behavior; whereas for socially unconfident girls and all boys, a social demonstration-avoid goal was negatively associated with aggressive behavior. This suggests that aggressive behavior for girls may have two different sources: desire for high social status, and concern about low social status. However, the latter results in aggressive behavior only if a girl views herself as socially skilled. The effect size for this interaction was small (.08), so it will be important for future research to replicate this finding. Future work that investigates relational aggression, in addition to overt aggression, might also provide additional information regarding gender differences in how social achievement goals manifest themselves in behavior among peers (Cillessen & Mayeux, 2004; Rose et al., 2004). It may be that gender differences in behaviors such as gossiping or rumor spreading are part of the explanation for the



increased aggressive behavior for socially skilled yet socially demonstration-avoidant girls.

Notably, the documented effects for social achievement goals were independent of perceived social competence. Thus, orientations to social competence explained variation in social adjustment above and beyond perceived social competence. As expected, approach social achievement goals had positive correlations with perceived social competence. Unexpectedly, a social demonstration-avoid goal had a small positive correlation with perceived social competence (although weaker than both approach goals). Perhaps a social demonstration-avoid goal is more related to other early adolescent issues (i.e., new social setting, puberty, and imaginary audience concerns) than perceived social competence. A social demonstration-avoid goal may, in part, reflect a strategic approach to navigating the social scene of a new school; that may be why there is a positive correlation with perceived social competence in our sample. In addition, correlations among goals revealed an interesting pattern for a social demonstration-avoid goal. Although positively correlated with a social demonstration-approach goal as expected, a social demonstration-avoid goal was also correlated with a social development goal. Although not unprecedented in the academic domain, this is less common. Thus, in general, the results indicate that the nature of achievement goals in the social domain is similar to other domains; however, there may be some differences in the nature of achievement goals between the social and academic domains.

### General Discussion

The present research establishes that young adolescents' different orientations toward developing or demonstrating social competence are an important element of social motivation in middle school. Although young adolescents are often concerned with validation and approval from peers (i.e., demonstration goals), they are also often oriented toward personal standards of developing themselves in their social world (i.e., development goals). Further, the distinction between garnering positive feedback from others and avoiding negative feedback from others is important (i.e., demonstration-approach vs. demonstration-avoid goals). Thus, an achievement goal approach is a facet of young adolescents' social motivation that has been generally overlooked.

We do not suggest that social achievement goals are the most critical or salient aspect of social motivation. They may be consequential because, as young adolescents accomplish or fail at the host of social goals that have been highlighted in previous research (intimacy, fun, etc.), they evaluate their social competence. The judgments early adolescents make of themselves and the processes that unfold are likely to be intimately related to their social achievement goals. Future work that examines the relations between other social goals that have been found to be important (e.g., intimacy, revenge, responsibility) and social achievement goals may further our understanding of social motivation during early adolescence.

Future research may also expand our understanding of social achievement goals, perceived social competence, and social adjustment by following students over a longer period of time and examining a wider array of adjustment indices. The implications of social achievement goals, as well as perceived social competence, may be different when considered in light of long-term develop-

ment. The nature and consequences of social achievement goals may vary at different stages of life. Although it seems that a social development goal would be adaptive and a social demonstration-avoid goal maladaptive at any age, a social demonstration-approach goal may not be associated with aggressive behavior at younger ages, when perhaps the peer culture is not as supportive of such behavior (LaFontana & Cillessen, 2002; Rose et al., 2004). In addition, the present research is correlational and cannot make conclusions about causality. Future work that examines causal relations, especially as affected by development, would be informative.

Further, future research with friend reports and peer nomination measures could provide another perspective on students' behaviors within school. Whereas students and teachers provide valid information regarding social adjustment, peers can provide insight into aspects of behavior that self and teacher reports cannot (e.g., whether relationships are reciprocated and how students' reputation stands among peers at school). For example, examining different measures of social status (sociometric popularity and peer perceptions of popularity) could be informative. Measures of sociometric popularity ask students about peers they like or do not like and thus indicate likeability. Measures of peer-perceived popularity ask students about who is popular or cool and thus indicate social status and centrality (Cillessen & Rose, 2005). Our results suggest that a social development goal would be associated with wide likeability, whereas a social demonstration-approach goal would be associated with a popular reputation.

As achievement goals have been shown to be influenced by features of the context (see Elliot, 2005, for a review), this line of research may have implications for creating middle school environments that promote social adjustment. Further, results from research concerning social achievement goals may be integrated with results concerning academic achievement goals and aid in making recommendations that simultaneously promote academic and social adjustment. As teachers are overwhelmed with numerous problems and issues, identifying key factors that might support multiple aspects of adjustment is important. Early adolescence marks the beginning of a downward trend in academic and social adjustment for many children (Juvonen et al., 2004). Researchers across many disciplines, as well as educators and parents, have become increasingly concerned about adjustment during early adolescence. The research reported here takes an initial step toward using a common framework (i.e., applicable to academic, athletic, and social domains) to better understand social adjustment during early adolescence with the hope that such an approach will ultimately contribute to knowledge that may help all students reach their potential in middle school.

### References

- Achenbach, T. M., McConaughtly, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, 101, 213-232.
- Aiken, L. S. & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage
- Anderman, L. H. (1999). Classroom goal orientation, school belonging and social goals as predictors of students' positive and negative affect following the transition to middle school. *Journal of Research & Development in Education*, 32, 89-103.



- Barron, K. E., & Harackiewicz, J. M. (2001). Achievement goals and optimal motivation: Testing multiple goal models. *Journal of Personality and Social Psychology*, 80, 706–722.
- Bouffard, T., Boisvet, J., Vezeau, C., & Larouche, C. (1995). The impact of goal orientation on self-regulation and performance among college students. *British Journal of Educational Psychology*, 65, 317–329.
- Cairns, R. B., Leung, M.-C., Gest, S. D., & Cairns, B. D. (1995). A brief method for assessing social development: Structure, reliability, stability, and developmental validity of the Interpersonal Competence Scale. *Behavioral Research and Therapy*, 33, 725–736.
- Cassidy, J., & Asher, S. R. (1992). Loneliness and peer relations in young children. *Child Development*, 63, 350–365.
- Chung, T., & Asher, S. (1996). Children's goals and strategies in peer conflict situations. *Merrill Palmer Quarterly*, 42, 125–147.
- Cillessen, A. H. N., & Mayeux, L. (2004). From censure to reinforcement: Developmental changes in the associations between aggression and social status. *Child Development*, 75, 147–163.
- Cillessen, A. H. N., & Rose, A. J. (2005). Understanding popularity in the peer system. *Current Directions in Psychological Science*, 14, 102–105.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46.
- Crick, N. R. (1996). The role of overt aggression, relational aggression, and prosocial behavior in the prediction of children's future social adjustment. *Child Development*, 67, 2317–2327.
- Duda, J. L., & Nicholls, J. G. (1992). Dimensions of achievement motivation in schoolwork and sport. *Journal of Educational Psychology*, 84, 290–299.
- Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist*, 41, 1040–1048.
- Dweck, C. S., & Leggett, E. L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review*, 95, 256–272.
- Eccles, J. (2004). Schools, achievement motivation and stage-environment fit. In R. M. Lerner & L. Steinberg (Eds.), *Handbook of adolescent psychology* (pp. 125–154). New York: Wiley.
- Eder, D. (1985). The cycle of popularity: Interpersonal relations among female adolescents. *Sociology of Education*, 58, 154–165.
- Eisenberg, N., & Morris, A. S. (2004). Moral cognitions and prosocial responding in adolescence. In R. M. Lerner & L. Steinberg (Eds.), *Handbook of adolescent psychology* (pp. 155–188). New York: Wiley.
- Elkind, D. (1967). Egocentrism in adolescence. *Child Development*, 38, 1025–1034.
- Elliot, A. J. (2005). A conceptual history of the achievement goal construct. In A. J. Elliot & C. Dweck (Eds.), *Handbook of competence and motivation* (pp. 52–72). New York: Guilford Press.
- Elliot, A. J., & Church, M. A. (1997). A hierarchical model of approach and avoidance achievement motivation. *Journal of Personality and Social Psychology*, 72, 218–232.
- Elliot, A. J., & Dweck, C. (2005). *Handbook of competence and motivation*. New York: Guilford Press.
- Elliot, A. J., Gable, S. L., & Mapes, R. R. (2006). Approach and avoidance motivation in the social domain. *Personality and Social Psychology Bulletin*, 32, 378–391.
- Elliot, A. J., & McGregor, H. A. (2001). A 2 × 2 achievement goal framework. *Journal of Educational Psychology*, 80, 501–519.
- Elliott, E. S., & Dweck, C. S. (1988). Goals: An approach to motivation and achievement. *Journal of Personality and Social Psychology*, 54, 5–12.
- Erdley, C. A., & Asher, S. R. (1996). Children's social goals and self-efficacy perceptions as influences on their responses to ambiguous provocation. *Child Development*, 67, 1329–1344.
- Erdley, C. A., Cain, K. M., Loomis, C. C., Dumas-Hines, F., & Dweck, C. (1997). Relations among children's social goals, implicit personality theories, and responses to social failure. *Developmental Psychology*, 33, 263–272.
- Ford, M. (1982). Social cognition and social competence in adolescence. *Developmental Psychology*, 18, 323–340.
- Ford, M. E. (1992). *Motivating humans: Goals, emotions, and personal agency beliefs*. Newbury Park, CA: Sage.
- Gazelle, H., & Ladd, G. W. (2003). Anxious solitude and peer exclusion: A diathesis-stress model of internalizing trajectories in childhood. *Child Development*, 74, 257–278.
- Gazelle, H., & Rudolph, K. D. (2004). Moving toward and away from the world: Social approach and avoidance trajectories in anxious solitary youth. *Child Development*, 75, 829–849.
- Grant, H., & Dweck, C. (2003). Clarifying achievement goals and their impact. *Journal of Personality and Social Psychology*, 85, 541–553.
- Harackiewicz, J. M., & Elliot, A. J. (1993). Achievement goals and intrinsic motivation. *Journal of Personality and Social Psychology*, 65, 904–915.
- Harter, S. (1982). The perceived competence scale for children. *Child Development*, 53, 87–97.
- Hawley, P. H. (2003). Prosocial and coercive configurations of resource control in early adolescence: A case for the well-adapted Machiavellian. *Merrill Palmer Quarterly*, 49, 279–309.
- Horst, S. J., Finney, S. J., & Barron, K. E. (2007). Moving beyond academic achievement measures: A study of social achievement goals. *Contemporary Educational Psychology*, 32, 667–698.
- Jarvinan, D. W., & Nicholls, J. G. (1996). Adolescents' social goals, beliefs about the causes of social success, and satisfaction in peer relations. *Developmental Psychology*, 32, 435–441.
- Juvonen, J., Le, V., Kaganoff, T., Augustine, C., & Constant, L. (2004). *Focus on the wonder years: Challenges facing the American middle school*. Santa Monica, CA: Rand.
- Kaplan, A., Gheen, M., & Midgley, C. (2002). Classroom goal structure and student disruptive behaviour. *British Journal of Educational Psychology*, 72, 191–212.
- Keating, D. P. (2004). Cognitive and brain development. In R. M. Lerner & L. Steinberg (Eds.), *Handbook of adolescent psychology* (pp. 45–84). New York: Wiley.
- Ladd, G. W., & Price, J. M. (1986). Promoting children's cognitive and social competence: The relation between parents' perceptions of task difficulty and children's perceived and actual competence. *Child Development*, 57, 446–460.
- LaFontana, K. M., & Cillessen, A. H. N. (2002). Children's perceptions of popular and unpopular peers: A multimethod assessment. *Developmental Psychology*, 38, 635–647.
- Lord, S., Eccles, J. S., & McCarthy, K. (1994). Risk and protective factors in the transition to junior high school students. *Journal of Early Adolescence*, 14, 162–199.
- McAdams, D. P. (1987). Motivation and friendship. In S. Duck & D. Perlman (Eds.), *Understanding personal relationships: An interdisciplinary approach* (pp. 85–105). Thousand Oaks, CA: Sage Publications.
- Middleton, M., & Midgley, C. (1997). Avoiding the demonstration of the lack of ability: An underexplored aspect of goal theory. *Journal of Educational Psychology*, 89, 710–718.
- Midgley, C., Kaplan, A., & Middleton, M. (2001). Performance-approach goals: Good for what, for whom, under what circumstances, and at what cost? *Journal of Educational Psychology*, 93, 77–86.
- Newcomb, A., & Bagwell, C. (1995). Children's friendship relations: A meta-analytic review. *Psychological Review*, 117, 306–347.
- O'Brien, S. F., & Bierman, K. L. (1988). Conceptions and perceived influence of peer groups: Interviews with adolescents and preadolescents. *Child Development*, 59, 1360–1365.
- Ojanen, T., Gronroos, M., & Salmivalli, C. (2005). An interpersonal circumplex model of children's social goals: Links with peer-reported behavior and sociometric status. *Developmental Psychology*, 41, 699–710.

- Parker, J. G., & Asher, S. R. (1993). Friendship and friendship quality in middle childhood: Links with peer group acceptance and feelings of loneliness and social dissatisfaction. *Developmental Psychology*, 29, 611–621.
- Parkhurst, J. T., & Hopmeyer, A. G. (1998). Sociometric popularity and peer perceived popularity: Two distinct dimensions of peer status. *Journal of Early Adolescence*, 18, 125–144.
- Pintrich, P. R. (2000). An achievement goal theory perspective on issues in motivation terminology, theory, and research. *Contemporary Educational Psychology*, 25, 92–104.
- Robins, R. W., & Pals, J. L. (2002). Implicit self-theories in the academic domain: Implications for goal orientation, attributions, affect, and self-esteem change. *Self and Identity*, 1, 313–336.
- Rodkin, P. C., Farmer, T. W., Pearl, R., & Van Acker, R. (2000). Heterogeneity of popular boys: Antisocial and prosocial configurations. *Developmental Psychology*, 36, 14–24.
- Rose, A. J. (2002). Co-rumination in the friendships of girls and boys. *Child Development*, 73, 1830–1843.
- Rose, A., & Asher, S. (1999). Children's goals and strategies in response to conflicts within a friendship. *Developmental Psychology*, 36, 69–79.
- Rose, A. J., Swenson, L. P., & Waller, E. M. (2004). Overt and relational aggression and perceived popularity: Developmental differences in current and prospective relations. *Developmental Psychology*, 40, 378–387.
- Ruble, D. N., Boggiano, A. K., Feldman, N. S., & Loebl, J. H. (1980). Developmental analysis of the role of social comparison in self-evaluation. *Developmental Psychology*, 16, 105–115.
- Rudolph, K. D., Caldwell, M. S., & Conley, C. S. (2005). Need for approval and children's well-being. *Child Development*, 76, 309–323.
- Ryan, A. M., Patrick, H., & Shim, S. O. (2005). Differential profiles of students identified by their teacher as having avoidant, appropriate or dependent help-seeking tendencies in the classroom. *Journal of Educational Psychology*, 97, 275–285.
- Ryan, A. M., & Shim, S. S. (2006). Social achievement goals: The nature and consequences of different orientations toward social competence. *Personality and Social Psychology Bulletin*, 32, 1246–1263.
- Savin-Williams, R. C., & Berndt, T. J. (1990). Friendship and peer relations. In S. S. Feldman & G. R. Elliott (Eds.), *At the threshold: The developing adolescent* (pp. 277–307). Cambridge, MA: Harvard University Press.
- Selman, R. (1980). *The growth of interpersonal understanding: Developmental and clinical analyses*. New York: Academic Press.
- Simmons, R., Rosenberg, F., & Rosenberg, M. (1973). Disturbance in the self-image at adolescence. *American Sociological Review*, 38, 553–568.
- Skaalvik, E. M. (1997). Self-enhancing and self-defeating ego orientation: Relations with task and avoidance orientation, achievement, self-perceptions, and anxiety. *Journal of Educational Psychology*, 89, 71–81.
- Smiley, P. A., & Dweck, C. S. (1994). Individual differences in achievement goals among young children. *Child Development*, 65, 1723–1743.
- Urdu, T. C. (1997). Examining the relations among early adolescent students' goals and friends' orientation toward effort and achievement in school. *Contemporary Educational Psychology*, 22, 165–191.
- Wentzel, K. (2000). What is it that I'm trying to achieve? Classroom goals from a content perspective. *Contemporary Educational Psychology*, 25, 105–115.
- Wentzel, K. (2001). The contribution of social goal setting to children's school adjustment. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 221–246). San Diego, CA: Academic Press.
- Wheeler, V. A., & Ladd, G. W. (1982). Assessment of children's self-efficacy for social interactions with peers. *Developmental Psychology*, 18, 795–805.
- Youniss, J., & Smollar, J. (1985). *Adolescent relations with mothers, fathers and friends*. Chicago: University of Chicago Press.

Received May 29, 2007

Revision received September 25, 2007

Accepted December 6, 2007 ■



# Students' Motivational Profiles and Achievement Outcomes in Physical Education: A Self-Determination Perspective

Julie C. S. Boiché and Philippe G. Sarrazin  
University of Grenoble

Frederick M. E. Grouzet  
University of Victoria

Luc G. Pelletier  
University of Ottawa

Julien P. Chanal  
University of Grenoble

Previous studies in education have inspected the relations between students' autonomous versus controlled motivation and relevant outcomes. In most of those studies a global index of self-determined motivation was created. The purpose of this article was to examine (a) how the different types of motivation proposed by Self-Determination Theory combine into distinct profiles as identified by cluster analysis and (b) the links between those profiles and objective criteria of achievement. In Study 1, motivation toward physical education was assessed at the beginning of a 10-week gymnastics teaching cycle, and performance was assessed at the end of the cycle among a sample of high school students ( $N = 210$ ). Study 2 ( $N = 215$ ) extended Study 1 by controlling students' initial performance, measuring the effort they exerted and recording their grades. Cluster analyses revealed three motivational profiles: self-determined, non-self-determined, and moderate levels of both types of motivation. Path analysis showed that the self-determined profile was related to the highest achievement. The results are discussed in terms of their implications for the assessment of students' motivation and the consequences of motivational profiles for educational outcomes.

**Keywords:** self-determination theory, motivational profile, achievement

The importance of motivation in education is unquestionable. As decades of research in educational settings have stressed, motivation is a consistent and significant contributor to students' functioning and performance (Good & Brophy, 2000). However, throughout the past 20 years, research using the framework of Self-Determination Theory (SDT; Deci & Ryan, 1985, 2000; Vallerand, 1997) has shown that individuals in general, and students in particular, differ considerably in the ways that they could be motivated toward an activity. More important, those differences in individuals' motivational orientations have far-reaching influences on their approach to an activity and the consequences that follow (see Deci & Ryan, 2000, for a review). This comprehensive framework holds the potential to contribute significantly to our

understanding of the issues related to motivation in education for the following reasons. First, it distinguishes between different types of motivation that can have a distinct impact on the maintenance and integration of behavior. Second, it presents clear hypotheses regarding the conditions that should hinder or facilitate students' motivation. Third, it outlines various consequences (cognitive, affective, and behavioral) that are associated with the different types of motivation (Vallerand, 1997). Fourth, it addresses the issue of internalization, the process by which behaviors that were initially reinforced by external sources (e.g., parents or teachers) become integrated within the individual to form a permanent part of his or her self.

In spite of the tremendous progress that has been made in this area of inquiry, some of the more basic questions concerning motivation conceptualization and measurement have remained relatively unexamined. One of the unexplored issues concerns the representation of the multiple forms of motivation proposed by SDT. Few studies have examined how these different goals combine to influence students' achievement behavior. In this article, we used cluster analysis in order to examine how the different forms of motivation proposed by SDT combine with each other and how they relate to students' academic performance in physical education (PE) classes. Two prospective studies were conducted in order to look at the links between the motivational profiles observed in a natural class setting and objective achievement criteria such as performance, overt effort, and grade. In the next part, we present the different motivational orientations assumed by SDT, as well as the links observed between those constructs and various academic outcomes. Then we explain how the different types of

---

Julie C. S. Boiché, Philippe G. Sarrazin, and Julien P. Chanal, Laboratoire Sport et Environnement Social, University of Grenoble, Grenoble, France; Frederick M. E. Grouzet, Department of Psychology, University of Victoria, Victoria, Canada; Luc G. Pelletier, School of Psychology, University of Ottawa, Ottawa, Ontario, Canada.

Julie C. S. Boiché is now at the Département STAPS, Faculté des Sciences de l'Homme et de l'Environnement, Université de la Réunion, Ile de la Réunion, France. Julien P. Chanal is now at the Faculté de Psychologie et des Sciences de l'Éducation, University of Genève, Genève, Switzerland.

We thank Aïna Chalabaev and Damien Tessier for their help in collecting the data and Robert Brustad for his thorough reading of the manuscript.

Correspondence concerning this article should be addressed to Philippe G. Sarrazin, University of Grenoble, Laboratoire Sport et Environnement Social, UFRAPS, BP53, 38041, Grenoble cedex 9, France. E-mail: philippe.sarrazin@ujf-grenoble.fr

motivation have been used in past literature and the problems that arise. Finally, we present the specific purposes of the two studies.

### The SDT Motivational Continuum

Research clearly supports the idea that individuals have different motivational orientations. They can be intrinsically motivated, when they are engaging in activities for their inherent satisfaction; extrinsically motivated, when they are engaging in activities for instrumental reasons; or amotivated, when they prove no regulation toward an activity. According to research, intrinsic motivation (IM) could derive from several sources. For instance, Vallerand (1997; Vallerand, Blais, Brière, & Pelletier, 1989) distinguished between IM to experience stimulation, when individuals are motivated by feeling pleasant sensations; IM toward knowledge, when they are moved by the desire to explore and learn new things; and IM toward accomplishment, when they aim at improving themselves.

Research has also supported distinctions regarding the different types of motivation that fall in the category of extrinsic motivation. The first two forms of extrinsic motivation are labeled, respectively, *external*, when the individuals' behavior is controlled by external sources, and *introjected*, when individuals have internalized the formerly external source of motivation but have not yet truly accepted the behavior. For these reasons, they are referred to as non-self-determined or controlled. The next two kinds of extrinsic motivation, on the other hand, represent self-determined, or autonomous, types of regulation. A distinction is made between identified regulation, which refers to a situation in which individuals perform an activity that has personal importance, and integrated regulation, in which individuals have integrated a behavior within their set of goals and values. Finally, the lowest level of self-determination proposed by SDT is amotivation. Amotivated individuals lack perceived competence because they do not feel able to perform the behavior, or they lack perceived control because they think their actions will not be adequate or sufficient to achieve a desired outcome (Deci & Ryan, 2000).

### The Outcomes of Motivation

A considerable amount of research has examined the relations between the different types of motivation and positive and negative outcomes. It has been observed that IM and self-determined forms of extrinsic motivation (i.e., identified and integrated regulations) have been linked to the more positive outcomes, whereas non-self-determined forms of motivation (i.e., external and introjected regulations) and amotivation have been linked to the less positive ones (Ryan & Deci, 2000; Vallerand, 1997). In education, there is evidence since the early 1990s that self-determined motivation toward school is related to several important outcomes (for reviews, see Deci, Vallerand, Pelletier, & Ryan, 1991; Reeve, 2002), including students' level of achievement (e.g., Burton, Lydon, D'Alessandro, & Koestner, 2006; Miserandino, 1996), coping style (e.g., Ryan & Connell, 1989), preference for optimal challenges (e.g., Boggiano, Main, & Katz, 1988), creativity (e.g., Amabile, 1985), well-being (Levesque, Zuehlke, Stanek, & Ryan, 2004), and persistence for a class (Vallerand & Bissonette, 1992) and for school (Vallerand, Fortier, & Guay, 1997).

Although research has been scarcer in the physical activity context, a growing body of research has confirmed this pattern of

results in that domain. Self-determined motivation has been linked to higher levels of self-reported effort in sport (Pelletier et al., 1995) and exercise (Fortier & Grenier, 1999) and to lower levels of sport dropout (Pelletier, Fortier, Vallerand, & Brière, 2001; Sarrazin, Vallerand, Guillet, Pelletier, & Cury, 2002). Also, self-determined motivation for physical education has been linked to the intention of being physically active or playing sports in the future (e.g., Ntoumanis, 2001; Standage, Duda, & Ntoumanis, 2003); to concentration, positive affect, and preference for challenging tasks (Standage, Duda, & Ntoumanis, 2005); to self-reported (Goudas, Biddle & Underwood, 1995; Ntoumanis, 2001, 2002) or teacher-reported effort (Ferrer-Caja & Weiss, 2000; Standage, Duda, & Ntoumanis, 2006); and to persistence and performance (Vansteenkiste, Simons, Lens, Sheldon, & Deci, 2004, Study 3); it has been negatively linked to boredom (Ntoumanis, 2001, 2002) or feelings of unhappiness (Standage et al., 2005).

### The Operationalization of Motivation

In order to test the SDT hypotheses, researchers have been faced with some crucial questions about the best ways to use the different types of motivation as determinants of various outcomes. Generally, correlational or experimental methods were used to examine relations between single goals and criterion measures (e.g., Ryan & Connell, 1989; Vansteenkiste et al., 2004). Sometimes, multiple regression analyses or structural equation modeling were used in order to test if goals have independent and additive effects for achieving a particular outcome (e.g., Ntoumanis, 2001; Pelletier et al., 2001). Finally, a major part of the studies integrated the scores obtained for the different types of motivations into a Self-Determination Index (e.g., Ryan & Connell, 1989). This index relies on an interactional hypothesis (Vallerand & Fortier, 1998), according to which intrinsic and extrinsic motivations are not independent, and a high level of one kind of regulation is necessarily linked to a low level of the other. The Self-Determination Index is generally calculated by giving each subscale a specific weight according to its respective place on the self-determination continuum (i.e., +3 +2, +1, -1, -2, and -3, respectively, for IM, integrated regulation, identified regulation, introjected regulation, external regulation, and amotivation scales). Next, the weighted scores of each subscale are added to derive a single index (e.g., Levesque et al., 2004; Miserandino, 1996; Vallerand et al., 1997). An important element in favor of the Self-Determination Index is the support for a matrix simplex in the continuum of self-determination. A matrix simplex is observed when the correlation between measures of two motivational constructs tends to decrease as the distance between them on the theoretical continuum increases. For example, because of its position on the theoretical continuum of self-determination, IM should be highly correlated with identified regulation but negatively correlated with amotivation. Its correlations with introjected and external regulation should take values comprised between the two others. This pattern of correlations has been observed in several domains, including education (e.g., Vallerand et al., 1993), sport (e.g., Pelletier et al., 1995), and exercise (Li, 1999).

However, several recent studies raised questions concerning the hypothesized motivational continuum, because the correlations among the ordered subscales provided only limited support for the simplex pattern (e.g., Cokley, 2000; Fairchild, Horst, Finney, &



Barron, 2005). For example, Fairchild et al. (2005) found that external regulation score is rather independent from the three intrinsic motivation scores (between .05 and .21). These results support the proposition that intrinsic and extrinsic motivation are not necessarily mutually exclusive, but rather independent constructs (e.g., Amabile, Hill, Hennessey, & Tighe, 1994; Covington & Mueller, 2001; Lepper & Henderlong, 2000). For example, Covington and Mueller (2001) underlined that "the weight of recent evidence suggests that intrinsic and extrinsic tendencies may best be conceived as two independent orientations, not just two endpoints on a single continuum" (p. 163). In this case, as Fairchild et al. (2005) noted, "perhaps one needs to consider how subscales combine or interact to promote motivation" (p. 335).

Little research has been carried out on the various ways to represent and group the different types of motivation proposed by SDT, in spite of the call of certain researchers to examine how they combine into distinct motivational profiles (Vallerand, 1997). This preoccupation is shared by other researchers, such as Sansone and Harackiewicz (2000), who concluded their book on intrinsic and extrinsic motivation by underlining that "the challenge that confronts theorists now is to specify *how* individuals might pursue more than one goal at a time and to detail the motivational dynamics of multiple goal pursuit" (p. 450). In the present article, we propose that a pattern-centered approach is suitable for situations in which several factors might act in conjunction with each other. It is especially the case when moderate or high correlations exist between several factors, which can undermine the efficacy of classical regression approaches (Mosteller & Tukey, 1977).

A particularly useful method of examining this issue is cluster analysis. This statistical technique identifies homogeneous groups, or clusters, on the basis of the shared characteristics they possess (Härdle & Simar, 2003). Therefore, the groupings obtained allow the researcher to examine differences between profiles rather than looking at interindividual differences. This kind of analysis should be helpful to determine if, in conformity with the SDT hypothesis, self-determined or non-self-determined profiles can be observed in natural settings, and/or to determine if, in conformity with the additive hypothesis, motivational profiles combining high levels of self-determined as well as non-self-determined motivation emerge.

Recently, several studies used cluster analysis to examine motivational profiles in the educational (e.g., Braten & Olaussen, 2005; Meece & Holt, 1993; Ntoumanis, 2002; Wang, Chatzisarantis, Spray, & Biddle, 2002), sport (Hodge & Petlichkoff, 2000; McNeill & Wang, 2005; Vlachopoulos, Karageorghis, & Terry, 2000), and physical activity settings (Biddle & Wang, 2003; Marshall, Biddle, Sallis, McKenzie, & Conway, 2002; Wang & Biddle, 2001), some of them having been based on achievement goal theory (e.g., Meece & Holt, 1993), SDT (Ntoumanis, 2002; Vlachopoulos et al., 2000), or a mix of both theories (e.g., Biddle & Wang, 2003; Wang et al., 2002). Two of these studies examined specifically the different motivations proposed by SDT. Vlachopoulos et al. (2000) found, among two large samples of athletes, one profile higher on self-determined forms of motivations and lower on non-self-determined forms of motivation, as well as one profile with relatively high scores on every kind of regulations and low scores on amotivation. None of the obtained profiles showed higher scores on non-self-determined than on self-determined forms of motivations. The researchers suggested that athletes with those profiles may have ceased their sport participation and therefore could not be part of the sample anymore.

The motivational profiles observed by Ntoumanis (2002) in physical education were quite different. In a cross-sectional study, he questioned two samples of British students about their experience in physical education. The questionnaire included self-reported measures of motivational climate, self-determined motivation, effort, enjoyment, and boredom. The data collected in the first school were used to conduct an exploratory cluster analysis, whereas the data from the second school were used to conduct a confirmatory analysis. The same three profiles emerged. In the first profile, students displayed high levels of self-determined motivation but low levels of external regulation and amotivation. In another cluster, students showed low scores for self-determined kinds of motivation, moderate scores for introjection, and high scores for external regulation and amotivation. The third profile was characterized by average scores for every form of regulation. These results have to be interpreted cautiously, however, because antecedent (motivational climate), consequences (e.g., boredom), and motivations were analyzed at the same time. In other words, the profiles observed in this study were not, strictly speaking, motivational profiles, because the numerous variables entered in the analysis may have influenced the results. It seems more appropriate to treat only the motivational scores with such an analysis, and afterwards to examine how the motivational profiles are related to certain important outcomes.

### The Current Studies

With the recent shift in interest for the possible ways to combine the different types of motivation proposed by SDT (i.e., the use of the Self-Determination Index vs. the use of different clusters), a key question concerns the relative predictive power of the different types of motivational profiles that could result from clusters analysis. That is, should motivation researchers who are interested in predicting important outcomes in education, such as involvement, performance, or grade, consider more than one way to combine the different forms of motivation proposed by SDT, and are some ways better suited to the prediction of some outcomes over others?

Accordingly, the purpose of this article was twofold. First, we examined whether the three profiles observed by Ntoumanis (2002) in a sample of British students would emerge among students from another country when only motivational variables in the grouping analysis were considered. The interest of this replication is both theoretical and empirical. From a theoretical perspective, it is not clear yet how the different kinds of motivation proposed by SDT should be combined. If a cluster analysis reveals only self-determined or not self-determined profiles of individuals, this would give support to the hypothesis of a continuum of motivation in the educational context. On the other hand, if cluster analysis reveals students' profiles with high levels of both self-determined and non-self-determined motivation, this would support the hypothesis that those two kinds of motivation can combine in naturalistic settings. Indeed, Lepper and Henderlong (2000) argued that "despite the experimental demonstrations that superfluous extrinsic contingencies *can* undermine intrinsic interest in controlled experimental contexts, [we think that] intrinsic and extrinsic motivation may, in many real-world settings, exert simultaneous positive influences on behavior" (p. 273). From an empirical point of view, there is great interest to know which motivational profiles actually exist in an academic setting like physical education classes and in which proportions students display such profiles.



In line with the findings of Ntoumanis (2002), two clusters coherent with the SDT hypothesis were expected to emerge, that is, where low levels of self-determined forms of motivation would be present with high levels of non-self-determined forms of motivations, and vice versa. We also expected the emergence of a third cluster that would show average levels of every kind of motivation. In order to reach this goal, an exploratory (Study 1) and a confirmatory (Study 2) cluster analysis were conducted on two separate samples of French high school students.

Second, we wanted to examine the consequences that those motivational profiles may have for behavioral variables representative of students' achievement (like effort, performance, and grade). It is important to emphasize that common limitations encountered in several of past studies conducted in education or in physical education were that most of the outcomes were assessed with self-reported measures and that these studies used a cross-sectional design (e.g., Ntoumanis, 2001, 2002; Standage et al., 2003, 2005). One problem with this procedure is that the links observed between self-reported measures from the same source may be inflated by shared variance rather than due exclusively to actual relations. Moreover, the absence of a longitudinal design limits the possibilities of inferring causality among the variables. We thus chose to use objective measures of students' achievement and to adopt a prospective design. In Study 1, we tested whether motivational profile would be linked to final performance, assuming that the more self-determined the profile of a student (i.e., high scores on self-determined scales and low scores on non-self-determined ones), the better would be his or her final performance. In Study 2, we adopted a more complex design in order to examine in more detail the role that motivational profile may play in the achievement process. Insofar as initial motivational profile might not be independent from initial performance, it seemed important to examine the relation between motivational profile and final performance while controlling for students' initial performance if we wanted to reinforce the idea that motivational profile actually has an effect on achievement. Grade and provided effort were also assessed as achievement outcomes.

## Study 1

### Method

#### Participants and Procedure

Two hundreds and ten students (104 girls, 104 boys; 2 students did not specify gender) from a French high school volunteered to participate in the study. They were in sixth-grade ( $N = 73$ ), seventh-grade ( $N = 70$ ), and ninth-grade ( $N = 67$ ) classes.<sup>1</sup> Their ages ranged between 10.7 and 16.8 years ( $M = 13.26$  years,  $SD = 1.49$ ). In France, physical education is a compulsory subject for all high school students. Generally, teachers teach different sports and physical activities in 10-week cycles (i.e., 10 lessons of 2 hr). The study was conducted during gymnastics cycles in scheduled physical education lessons. Prior to the initiation of the research, teachers, parents, students, and school administrators were asked to participate in an observational study. Students' motivations for the activity were assessed during the first lesson. At the end of the cycle, all students were videotaped individually in order to evaluate their performance in gymnastics.

### Measures

**Motivation.** Motivation toward gymnastics was assessed at the beginning of the cycle. An adaptation of French motivation scales in sport (i.e., the "Echelle de Motivations dans les Sports" [EMS], Brière, Vallerand, Blais, & Pelletier, 1995) and in education (i.e., the "Echelle de Motivation Académique," Vallerand et al., 1989), and an adaptation of a scale developed in English to assess physical education motivation (Standage et al., 2006), were used in order to fit both the sportive and educational aspects of physical education. These three tools assessed the multifaceted motivational regulations proposed by SDT. Depending on the subscale, the items were preferentially adapted from one tool or the other. For example, the IM scales were mainly derived from the EMS. The adaptation consisted of minor changes in the wording of some items to target the gymnastics context and/or a translation in French. Six motivational constructs relative to physical education were assessed in an 18-item scale. The participants had to complete the following sentence: "I participate in gymnastics classes" with items reflecting IM to experience stimulation (e.g., "for the excitement I feel when I am really involved in the activity"), IM toward knowledge or accomplishment (e.g., "for the satisfaction I experience while I am perfecting my abilities"),<sup>2</sup> identified regulation (e.g., "because what I learn in this activity will be useful later"), introjected regulation (e.g., "because I would feel guilty if I could not succeed in this activity"), external regulation (e.g., "because that's what I'm supposed to do"), and amotivation (e.g., "I don't know why I go in gymnastics, if I could, I would get exempted"). An English version of the entire scale is presented in Table 1. Responses were made on a 7-point scale ranging from 1 (*strongly disagree*) to 7 (*strongly agree*).

**Performance in gymnastics.** Students' performance in gymnastics was based on a test consisting of five basic gymnastics exercises (e.g., stretched handstand back drop, backward roll, and cartwheel). The students were videotaped one by one during a class prior to the initiation of the gymnastics cycle. Three experts in gymnastics rated the videotaped performance of each student on

<sup>1</sup> In France, high school is comprised of junior high school, which usually includes students who are 11 to 15 years old and between sixth and ninth grades, and senior high school, which includes older students between tenth and twelfth grade. Both of our samples consisted of junior high school students (i.e., those in the four first grades of high school).

<sup>2</sup> Few authors (e.g., Cokley, Bernard, Cunningham, & Motoike, 2001; Fairchild et al., 2005) questioned the distinctiveness of IM toward knowledge and accomplishment subscales, insofar as they appeared very correlated (.86). Results of a pilot study that we carried out with a representative sample of 100 students showed that scores on IM toward knowledge and accomplishment were very correlated ( $\phi = .92$ ). A test of the discriminant validity between the two constructs, which consisted of examining whether the pair of latent factors could be treated as a single construct by setting each correlation to 1.0 and comparing the constrained model with a model in which the correlation was free to vary, revealed that correlation was not significantly different from 1.0 and thereby did not support the discriminant validity of the constructs. Finally, interviews carried out with 10 students of this sample showed that they did not perceive the nuances that could exist between items assessing IM toward knowledge (e.g., "For the pleasure I experience when I learn new skills") and those assessing IM toward accomplishment (e.g., "For the pleasure I experience when I improve some of my weaknesses"). As a result, only three items among the most representative of those subscales have been retained in this study.



Table 1  
*Items of the Motivation Scale With Corresponding Factor Loadings (Study 1): "I participate in gymnastics classes . . ."*

Items	Loadings
IM Toward Stimulation	
For the excitement I feel when I am really involved in the activity	.82
Because this activity is fun	.93
For the emotions I experience while I practice this activity	.87
IM Toward Knowledge or Accomplishment	
For the pleasure I experience when I improve some of my weaknesses	.75
For the pleasure I experience when I learn new skills	.88
For the satisfaction I experience while I am perfecting my abilities	.91
Identified Regulation	
Because what I learn in this activity will be useful later	.80
Because it is important and it can help me for other things	.88
Because this activity is important for my personal growth	.71
Introjected Regulation	
Because I absolutely need to succeed in this activity	.77
Because I would feel bad about myself if I did not	.78
Because I would feel guilty if I could not succeed in this activity	.62
External Regulation	
Because I don't want to disappoint important people	.37
Because I'll get into trouble if I don't do what the teacher's asking	.56
Because that's what I'm supposed to do	.72
Amotivation	
I don't know why I go in gymnastics, if I could, I would get exempted	.69
But it doesn't worth it, I feel that I'm wasting my time	.74
I don't see why we should have gymnastics	.77

*Note.* This table presents an English translation of the French items used in the studies.

a 1 (*low level*) to 7 (*high level*) response scale. Each expert looked at the videotapes individually and made a single global rating of the student's performance after viewing all five exercises. Interjudge reliability was very good ( $\alpha > .90$ ). The mean of the three experts' scores was calculated and used as an indication of students' gymnastics performance. This procedure has been used previously in studies involving physical education classes (e.g., Chanal, Marsh, Sarrazin, & Bois, 2005).

### Data Analysis

First, we carried out a confirmatory factor analysis (using Lisrel 8.54; Jöreskog & Sörbom, 2003) in order to examine the factorial structure of the questionnaire assessing students' motivation. Next, the correlations between motivational scores were examined. Afterwards, an exploratory cluster analysis was conducted in order to examine how many motivational profiles could be identified. An examination of the cluster composition according to gender and school year was then carried out. Finally, an analysis of variance (ANOVA) was done in order to test whether motivational profiles were linked to final performance.

### Results and Discussion

The means and standard deviations of the variables are shown in Table 2, as well as the alpha coefficients and correlations among variables.

#### Confirmatory Factor Analysis

The structure we tested was supposed to rely on six factors, namely the following: IM toward stimulation, IM toward knowledge or accomplishment, identified regulation, introjected regulation, external regulation, and amotivation, with three items expected to load on each factor. A test carried out in order to evaluate the normality of the distribution of the variables revealed that none of them was normally distributed (multivariate skewness and kurtosis tests,  $ps < .0001$ ). We thus chose to use the robust maximum likelihood estimation procedure. The analysis showed a relatively good fit of the six-factor structure with the data,  $\chi^2(120) = 304$ ,  $p < .01$ , robust comparative fit index (CFI) = .97, non-normed fit index (NNFI) = .96, standardized root mean residual (SRMR) = .06, root-mean-square error of approximation (RMSEA) = .07, 90% confidence interval (CI) of RMSEA = .057–.082. The factor loadings were all significant at the

Table 2  
Descriptive Statistics and Correlations Between the Variables (Study 1)

Variable	<i>M</i>	<i>SD</i>	$\alpha$	2	3	4	5	6	7
1. IM toward stimulation	3.56	1.99	.90	.65****	.63****	.52****	-.32****	-.52****	.26****
2. IM toward knowledge and accomplishment	4.15	1.92	.88		.75****	.62****	-.28****	-.53****	.17**
3. Identified regulation	3.29	1.80	.84			.59***	-.19***	-.50****	.13*
4. Introjected regulation	3.24	1.60	.69				-.02	-.25****	.01
5. External regulation	3.51	1.71	.67					.57****	-.14**
6. Amotivation	3.28	1.95	.78						-.37****
7. Final performance in gymnastics	4.14	1.70	.93						

Note. IM = intrinsic motivation.

\*\*\*\*  $p < .001$ . \*\*\*  $p < .01$ . \*\*  $p < .05$ . \*  $p < .07$ .

.01 level and ranged between .37 and .93 ( $M = .74$ ). They are presented along with the corresponding items in Table 1.

### Simplex Pattern

Correlations among the constructs are shown in Table 2. The correlations between the variables appear to be in conformity with a simplex ordered matrix, although we found some deviations from this presumed pattern. For example, IM toward knowledge or accomplishment displayed a more important relationship with identified regulation (.75) than with IM toward stimulation (.65). More important, introjected regulation showed strong positive correlations with IM toward stimulation (.52), IM toward knowledge or accomplishment (.62) and identified regulation (.59), no correlation with external regulation ( $-.02$ , *ns*), and only moderate negative correlations with amotivation ( $-.25$ ).

### Cluster Analysis

Given that the internal consistency of each subscale proved to be satisfactory (see Table 2), the average of the scores was calculated for each one. A cluster analysis was conducted next, following the procedure suggested by Hair, Anderson, Tatham, and Black (1998). All the variables included in a cluster analysis have to share the same metric so that each of them contributes equally to the formation of the clusters. All the motivational constructs were assessed on a 7-point

scale. Cluster analysis is also sensitive to outliers. Preliminary analyses showed no cases with a distance from the mean greater than three times the value of the standard deviation. Finally, multicollinearity between variables may impact on the cluster analysis by giving more weight to collinear variables. Given that no Bravais-Pearson correlation coefficient was higher than .90, we considered that there was no problem of this kind (Hair et al., 1998).

A hierarchical cluster analysis was performed using Ward's method with a squared Euclidean distance measure. The agglomeration schedule and dendrogram were used to identify the number of clusters. A high increase of the agglomeration schedule (37%) suggested a three-cluster solution to be suitable. Figure 1 shows the profiles of those three clusters. The first cluster was labeled "self-determined" profile and represented 35% of the sample ( $N = 73$ ). Students in this cluster showed high levels of self-determined forms of motivation (i.e., IM and identified regulation), moderate level of introjected regulation, and low levels of external regulation and amotivation. The second cluster was labeled "moderate" profile and also represented 35% of the sample ( $N = 73$ ). Students in this cluster had average scores for every form of motivation, with a majority of scores close to 3 and 4 on a 7-point scale. The third cluster was labeled "non-self-determined" profile and represented 30% of the sample ( $N = 64$ ). Students in this cluster displayed low levels of self-determined forms of motivation (i.e., IM and identified regulation), a low level of introjected regulation, and relatively high levels of external regulation and amotivation.

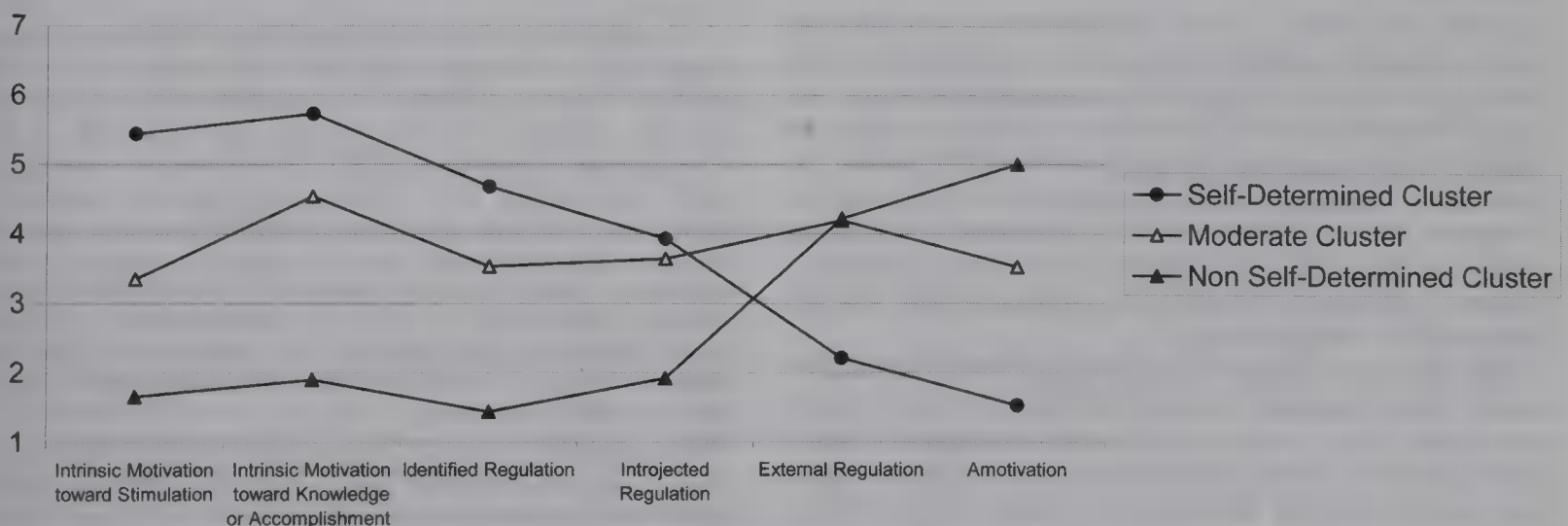


Figure 1. Motivational profiles in Study 1.



Table 3  
Descriptive Statistics for the Three Clusters (Study 1)

Variable	Cluster 1 ( <i>N</i> = 73), self-determined		Cluster 2 ( <i>N</i> = 73), moderate		Cluster 3 ( <i>N</i> = 64), non- self-determined		<i>F</i>	<i>p</i>	$\eta^2$
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			
1. IM toward stimulation	5.44 <sup>a</sup>	1.29	3.35 <sup>b</sup>	1.43	1.66 <sup>c</sup>	1.03	212.11	.001	.65
2. IM toward knowledge or accomplishment	5.73 <sup>a</sup>	1.25	4.54 <sup>b</sup>	1.10	1.90 <sup>c</sup>	0.91	218.11	.001	.66
3. Identified regulation	4.68 <sup>a</sup>	1.59	3.53 <sup>b</sup>	1.21	1.44 <sup>c</sup>	0.59	131.65	.001	.54
4. Introjected regulation	3.93 <sup>a</sup>	1.53	3.64 <sup>a</sup>	1.37	1.92 <sup>b</sup>	1.05	46.82	.001	.29
5. External regulation	2.21 <sup>a</sup>	1.25	4.19 <sup>b</sup>	1.49	4.22 <sup>b</sup>	1.60	96.15	.001	.46
6. Amotivation	1.53 <sup>a</sup>	1.77	3.52 <sup>b</sup>	1.33	5.00 <sup>c</sup>	1.82	202.71	.001	.64
7. Final performance in gymnastics	4.70 <sup>a</sup>	1.60	3.93 <sup>b</sup>	1.63	3.51 <sup>c</sup>	1.71	5.37	.01	.09

Note. Means in the same row that do not share superscripts differ at  $p < .01$ , using Newman-Keuls post hoc tests. IM = intrinsic motivation.

The clusters' size, as well as the means and standard deviations of their centroid, are shown in Table 3.

A multivariate analysis of variance (MANOVA) showed a significant effect of cluster membership on the six motivational constructs (see Table 3; Wilks's lambda = .13, Rao  $R[12, 404] = 58.29$ ,  $p < .001$ ). Follow-up ANOVAs revealed a significant effect of cluster membership on each motivational construct (the values of  $F$ ,  $p$ , and  $\eta^2$  are indicated in Table 3). Newman-Keuls post hoc analyses ( $p < .01$ ) indicated that the three groups were significantly distinct from each other on all motivational scales, except for introjected regulation between the moderate and self-determined profiles and for external regulation between the non-self-determined and moderate profiles.<sup>3</sup>

### Motivational Profile and Gymnastics Achievement

An ANOVA was carried out in order to test whether motivational profile was linked to different levels of achievement among students, represented by their final performance in the activity. The analysis was significant,  $F(2, 112) = 5.37$ ,  $p < .01$ ,  $\eta^2 = .09$ . Newman-Keuls post hoc analyses ( $p < .01$ ) revealed that the three clusters were distinct on final performance in gymnastics (see Table 3), with students in the self-determined cluster showing the highest average performance ( $M = 4.70$ ), followed by the moderate ( $M = 3.93$ ) and the non-self-determined cluster ( $M = 3.51$ ).<sup>4</sup>

Globally, this result is coherent with the tenets of SDT and with previous research in education (e.g., Reeve, 2002). However, the result relative to introjected regulation is somewhat surprising. As a non-self-determined form of regulation, it is theoretically assumed to lead to negative consequences. In the present study, it was found to be linked to better achievement. In the sport context, introjection has been found previously to be related to short term persistence (Pelletier et al., 2001) but not to long-term persistence. Perhaps it deserves to be considered a potential positive form of motivation at a short-term level.

This study also revealed the existence of a third motivational profile, which comprises students with moderate scores for all motivational scales. It thus seems possible to combine both self-determined and non-self-determined forms of motivation when they remain moderately developed. The analysis did not show a profile with high scores on the majority of subscales. The analyses showed that students with this moderate profile globally realized

better performances than did students with a non-self-determined profile, but poorer performances than students with a self-determined profile.

Even if a prospective design was adopted in this study, the correlational nature of the data does not allow us to affirm that the different motivational profiles observed were the cause for different levels of performance at the end of the cycle. In other words, we cannot ignore the possibility that differences in initial performance are responsible for the motivational profiles observed. It thus seemed important to control for initial performance in order to evaluate more accurately the potential impact of motivation on final performance. Moreover, criteria other than performance should also be used as representative of achievement.

### Study 2

The first aim of Study 2 was to verify that the motivational profiles observed in Study 1 could be validated in another sample of students. The second aim was to test the effect that those motivational profiles may have on performance, controlling for initial performance, as well as the role that these variables and exerted effort may play in the prediction of final grade.

According to SDT, it was hypothesized that the more self-determined the profile of the students, the greater their effort,

<sup>3</sup> Chi-squared tests of association were conducted in both studies to examine the possible link between cluster membership and gender or school year. The analyses revealed no association between students' gender and cluster membership. On the other hand, school year tended to be related to cluster membership in Study 1 (students from sixth grade being slightly overrepresented in the self-determined cluster and underrepresented in the non-self-determined cluster; an opposite pattern was observed for students from ninth grade) and was significantly related to cluster membership in Study 2 (students from eighth grade being slightly overrepresented in the moderate cluster and underrepresented in the other clusters). These results are coherent with past literature in education that indicated a decrease of self-determined motivation as age increases (Otis, Grouzet, & Pelletier, 2005; Ratelle, Guay, Larose, & Senécal, 2004).

<sup>4</sup> An additional 3 (profile)  $\times$  3 (grade) ANOVA was conducted. It revealed (a) a significant effect of grade ( $p < .05$ ), the performance tended to be higher with the grade level; (b) a significant effect of the motivational profile ( $p < .01$ ), with a significant difference between the three profiles ( $p < .05$ ); and (c) no interaction effect ( $p = .28$ ).

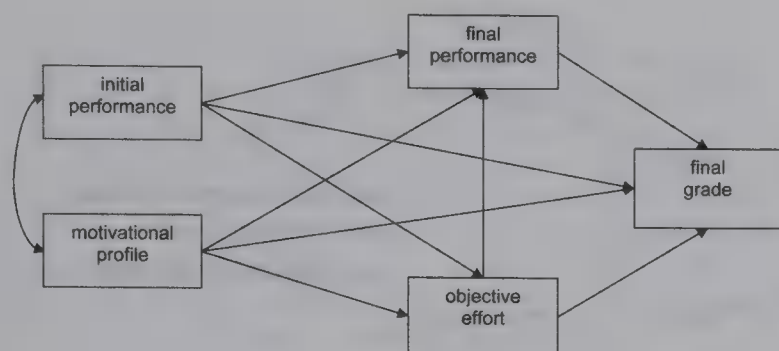


Figure 2. Hypothesized model of achievement.

performance, and grade would be. More precisely, the best consequences were expected to be achieved by students displaying a self-determined profile, followed by students showing a moderate profile. Students with a non-self-determined profile were expected to show the lowest level of achievement. Our hypothesized model of achievement (see Figure 2) comprised three dependent variables. The effort exerted by the students was expected to be related to both their initial performance in the activity and to their motivational profile. Their final performance was expected to be related to their initial performance, their motivational profile, and their effort. In other words, for a similar initial performance in the activity, we expected that putting more effort and being more self-determined for the activity would result in higher improvement, and thus in a higher final performance. Finally, the grade obtained was hypothesized to depend not only on students' final performance, but also on initial and middle cycle variables, that is, their initial performance and motivation as well as the objective effort they exerted. Teachers, and particularly physical education teachers, have been shown to prefer hardworking students, whatever their level of ability (Biddle & Goudas, 1997; Covington & Omelich, 1979; Weiner, 1979). Moreover, teachers might take into account initial characteristics of their students (e.g., performance and motivation) to form expectations on their achievement, which may impact on the grade they give them at the end of the cycle (e.g., Jussim & Eccles, 1992; Trouilloud, Sarrazin, Martinek, & Guillet, 2002).

## Method

### Participants and Procedure

Two hundreds and fifteen students (99 girls, 116 boys) from a French high school volunteered to participate in the study. They were in seventh-grade ( $N = 115$ ), eighth-grade ( $N = 70$ ), and ninth-grade ( $N = 30$ ) classes. Their ages ranged between 10.5 and 16 years ( $M = 12.47$  years,  $SD = 1.05$ ), and all were involved in a gymnastics cycle. During the first lesson, students' motivation and their initial performance in the activity were assessed. Next, students' effort in gymnastics was measured in the middle of the cycle (i.e., between the fourth and sixth gymnastics courses). Finally, students' performance in gymnastics was appraised during the last class, and their grade was recorded.

### Measures

**Motivation.** The same scale as in Study 1 was used.

**Performance in gymnastics.** The assessment of students' performance in the activity at the beginning and at the end of the teaching cycle was identical to Study 1.

**Effort in gymnastics.** In the middle of the cycle, a videotape was used to evaluate the amount of effort provided by the students. The camcorder was placed in a fixed spot, in order to record the activity of the students when they worked on the stretched handstand back-drop task. Several students were involved in this workshop simultaneously, but there was enough space so that they could work at it at the same time. Furthermore, teachers were previously informed of the purpose of the camcorder and were asked not to interact with students completing this exercise. The number of repetitions of the exercise for 5 min was calculated and used as an indication of their investment in the activity.

**Grade.** After completion of the gymnastics cycle, all teachers provided the grades obtained by the students in gymnastics. French grades are scores comprised between 0 and 20, and the score of 10 is usually considered a mediocre result.

### Data Analysis

A confirmatory cluster analysis was conducted on the sample, taking the number of clusters emerging in Study 1 as a basis for the

Table 4  
Descriptive Statistics for the Three Clusters (Study 2)

Variable	<i>M</i>	<i>SD</i>	$\alpha$	Cluster 1 ( $N = 84$ ), self-determined		Cluster 2 ( $N = 91$ ), moderate		Cluster 3 ( $N = 40$ ), non- self-determined	
				<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
1. IM toward stimulation	3.56	1.99	.88	5.67	1.03	3.09	1.11	1.71	0.80
2. IM toward knowledge or accomplishment	4.15	1.92	.84	5.65	1.14	3.98	1.13	1.94	0.90
3. Identified regulation	3.29	1.80	.80	5.13	1.29	3.22	1.07	2.06	1.08
4. Introjected regulation	3.24	1.60	.67	4.23	1.39	3.39	1.20	1.90	0.92
5. External regulation	3.51	1.71	.72	2.05	1.40	2.83	1.39	5.11	1.35
6. Amotivation	3.28	1.95	.88	1.77	1.30	2.80	1.31	5.94	1.14
7. Provided effort	4.35	3.43		4.72	2.68	4.46	3.60	3.17	2.54
8. Initial performance in gymnastics	3.32	1.10	.91	3.68	1.08	3.22	1.12	2.76	1.41
9. Final performance in gymnastics	4.65	1.48	.92	5.25	1.40	4.49	1.42	3.68	1.58
10. Grade	12.62	3.65		14	3.47	12.44	3.39	9.91	4.05

Note. IM = intrinsic motivation.



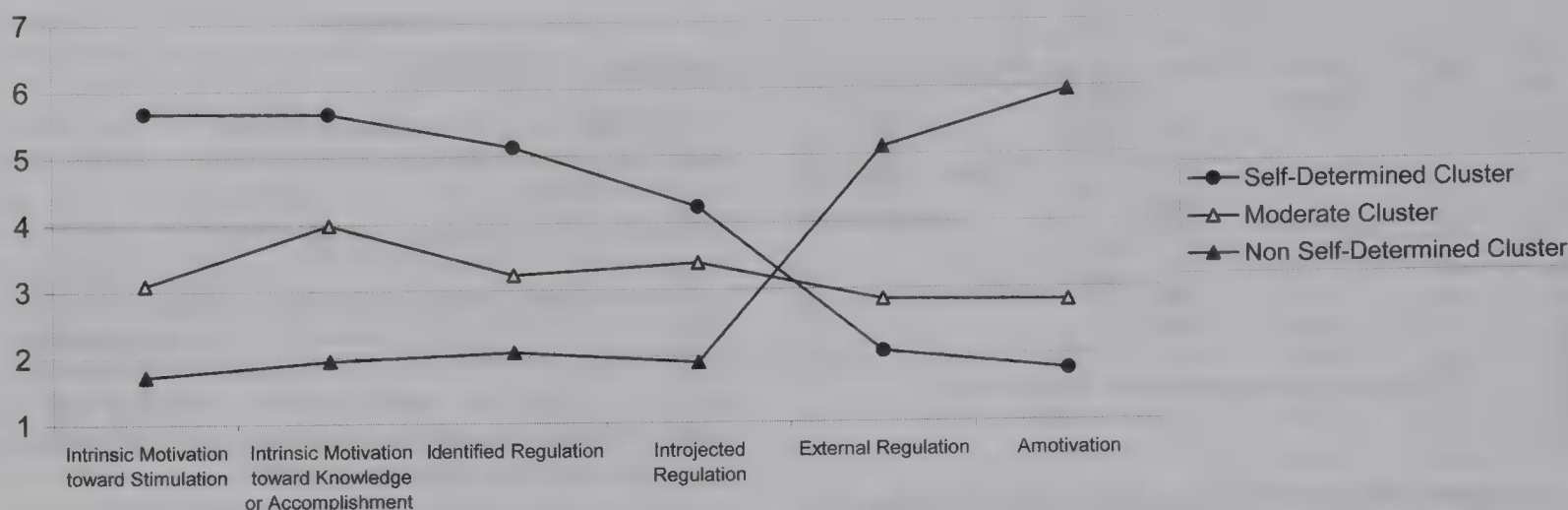


Figure 3. Motivational profiles in Study 2.

analysis. Next, we tested the hypothesized model of achievement (see Figure 2) using path analysis.

### Results and Discussion

The means and standard deviations of the variables are shown in Table 4, as well as the alpha coefficients for multi-item variables and Bravais-Pearson correlations among all variables.

#### Confirmatory Cluster Analysis

A k-mean cluster analysis was conducted on the basis of the result of the hierarchical cluster analysis carried out in Study 1. This kind of analysis is considered confirmatory because it requires the researcher to provide a specific number of clusters expected to emerge in the sample. In line with the result obtained with the first sample, the fact that the second sample was comparable in terms of age and gender and that it originated from a school presenting relatively similar characteristics, the number specified in the cluster analysis was set at three. A MANOVA showed a significant multivariate main effect of cluster membership on motivational scores, Wilks's lambda = .12, Rao  $R(12, 420) = 64.66$ ,  $p < .01$ . Follow-up ANOVAs showed that cluster membership led to a significant effect on each motivational construct ( $ps < .001$ ,  $\eta^2 > .29$ ). Newman-Keuls post hoc analyses ( $p < .01$ ) revealed that the three groups were significantly distinct from each other on all motivational scales. The motivational profiles were similar to those observed with the hierarchical procedure in Study 1 (see Figure 3), with little change regarding the repartition of the students in the various clusters. The self-determined profile represented 39% of the sample ( $N = 84$ ), the moderate profile 42% ( $N = 91$ ), and the non-self-determined profile 19% ( $N = 40$ ). Students in the non-self-determined profile demonstrated higher scores for external regulation and amotivation than did those in Study 1. The clusters' size, as well as the means and standard deviations of their centroid, are shown in Table 4.

#### Motivational Profiles and Achievement Outcomes

The hypothesized model (see Figure 2) was tested with path analysis using Lisrel 8.54 (Jöreskog & Sörbom, 2003). Two contrasts were computed in order to compare the motivational profiles.

The use of contrasts is considered particularly appropriate in situations where specific hypotheses are formulated by the researcher concerning between groups differences on a dependent variable (Cohen, Cohen, West, & Aiken, 2003). We expected students with a self-determined profile to show the best level of achievement, followed by students with a moderate profile, and finally students with a non-self-determined profile. In order to compare students with a non-self-determined versus moderate profile, the weights used to compute the first contrast were  $-1, 0$ , and  $1$  for the non-self-determined, self-determined, and moderate profiles, respectively. The second contrast compared students with a moderate versus self-determined profile. The weights were  $-1, 0$ , and  $1$  for the moderate, non-self-determined, and self-determined profiles, respectively. A covariance matrix was used with the maximum likelihood estimation method. Because the model we tested was saturated, no adjustment fit indices are given. Only the weights of the paths as well as their level of significance are indicated in Figure 4.

The analysis revealed that final performance was predicted by the motivational profiles, controlling for initial performance ( $\beta = .62$ ). Students with a moderate profile obtained better performances than did those with a non-self-determined profile ( $\beta = .27$ ), and students with a self-determined profile obtained better performances than did those with a moderate profile ( $\beta = .30$ ).<sup>5</sup> The two contrasts and initial performance explained 52% of final performance's variance. On the other hand, students' final grade was predicted by their final performance ( $\beta = .44$ ), their objective effort ( $\beta = .13$ ), and their motivational profile, while controlling for their initial performance in gymnastics ( $\beta = .27$ ). More specifically, as was the case for final performance, students with a moderate profile obtained better grades than did those with a non-self-determined profile ( $\beta = .23$ ), and students with a self-determined profile obtained better grades than did those with a moderate profile ( $\beta = .25$ ). Altogether those variables explained 59% of students' grade variance. Objective effort was predicted by

<sup>5</sup> When a contrast shows a significant link with an outcome, it means that the two profiles differ significantly on this variable. If both of the contrasts show significant links, then the three profiles are distinct from each other, because if self-determined  $\neq$  moderate, and moderate  $\neq$  self-determined, then non-self-determined  $\neq$  self-determined.

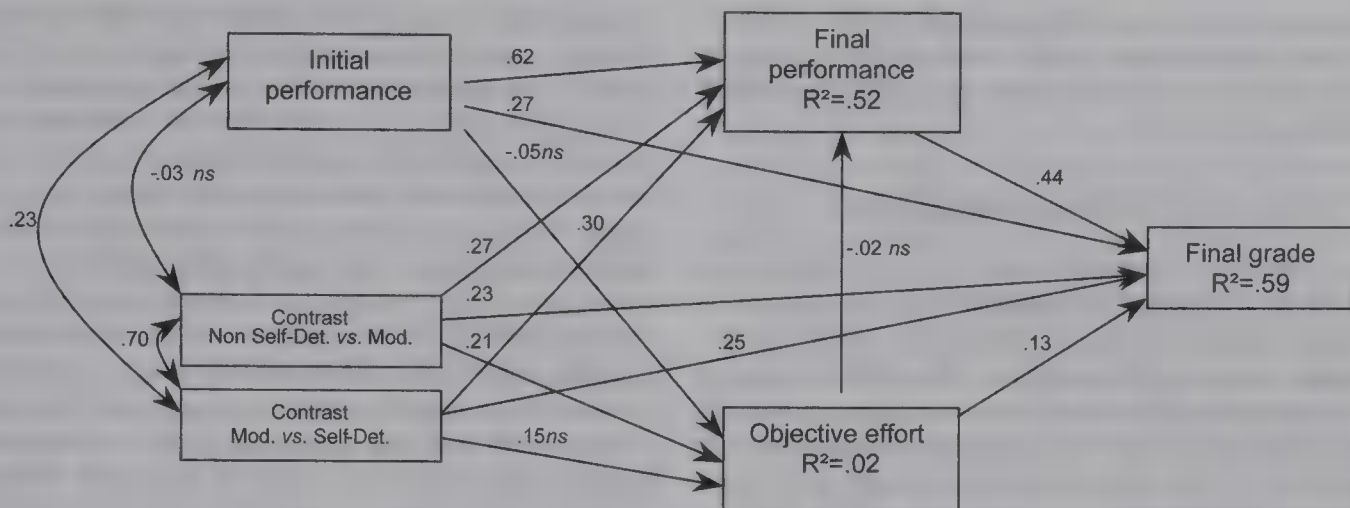


Figure 4. Path analysis contrasting the self-determined, moderate, and non-self-determined profiles (all paths are significant at  $p < .05$  except when specified by *ns*).

students' motivational profile while controlling for initial performance ( $\beta = -.05$ , *ns*). Students with a non-self-determined profile exerted more effort than did those with a non-self-determined profile ( $\beta = .21$ ), whereas there was no significant difference between students with a moderate or self-determined profile ( $\beta = .15$ , *ns*).<sup>6</sup>

Taken together, the analyses showed that students who showed a self-determined profile were those who achieved the most during the cycle: They provided more effort, realized the best final performance, and obtained better grades than did the students showing alternative profiles. By opposition, students who showed a non-self-determined profile were those who achieved less. A moderate motivational profile with average scores on all subscales was found to be linked to in-between achievement scores. One could assume that students showing this type of motivational profile are willing to provide effort in some certain circumstances, for example, if self-consciousness is salient. This might have been the case in this study, as the camcorder was visible, and thus some students may have tried to do the exercise just because this effort would be noticed. This could explain why no significant difference was observed between the moderate profile and the self-determined profile for this variable. However, this "superficial" effort does not seem sufficient to make progress, because there was actually a significant difference in terms of final performance between those two groups even after controlling for initial performance in the activity.

### General Discussion

Several authors have recently called on researchers to begin to pay attention to the possible ways to combine different types of motivation, and more specifically, the different types of motivation proposed by SDT (Fairchild et al., 2005; Vallerand, 1997). The purpose of this article was to adopt an original approach to examine students' motivation and the potential consequences that conceptualizing and regrouping different forms of motivation may have in education. More specifically, we chose to adopt a pattern-centered approach instead of testing specific effects of each motivational construct from SDT, or using an aggregated index of self-determination. That is, we examined how the different types of motivation combined into certain profiles and how those profiles were related to students' achievement. On one hand, recent

research has suggested that the Self-Determination Index may not be the only way to represent the different types of motivation proposed by SDT (e.g., Fairchild et al., 2005). On the other hand, looking at the separate effects of the different forms of motivation does not take into account the relationships between those motivations. Two prospective studies were conducted in a natural physical education setting in order to examine those issues.

### Description of the Motivational Profiles

Cluster analyses were conducted to examine which motivational profiles would emerge in our samples. Overall, the results confirmed those observed by Ntoumanis (2002) with British students, even if the procedure we adopted was not strictly identical. We observed two profiles with a clear motivational orientation toward either extremity of the theoretical continuum proposed by SDT (e.g., Deci & Ryan, 2000). These profiles provided support for the interactional hypothesis about the relations between motivational constructs (e.g., Vallerand & Fortier, 1998). The first cluster was oriented toward the self-determined part of the continuum of motivation, with high levels of IM and identified regulation, moderate levels of introjected regulation, and low levels of external regulation and amotivation. A second cluster was oriented toward the non-self-determined part of the continuum, with students scoring high on external regulation and amotivation, and low on other forms of motivation. The third motivational profile we observed presented average scores on each motivational variable. As proposed by some researchers (e.g., Lepper & Henderlong, 2000), it is possible that intrinsic and extrinsic forms of motivation combine in real-world settings. The results observed within our samples suggest that this can be the case if the level of those motivations remains moderate (see Figure 3). Compared with the results of Ntoumanis (2002), the repartition of the students in the clusters was rather similar for the moderate motivational profile, but fewer

<sup>6</sup> An additional model in which students' age was used as covariant revealed the same pattern of results and no significant association between this variable and the three objective criteria of students' achievement ( $\beta = .01$ ,  $p > .89$ ;  $\beta = -.06$ ,  $p > .25$ ;  $\beta = -.01$ ,  $p > .88$ , for effort, final performance, and final grade, respectively).



French students displayed a self-determined profile and more showed a non-self-determined profile. These differences can be explained by the fact physical education is compulsory for all high school students in France.

### *Motivational Profile and Achievement*

The second purpose of the studies was to examine the relations between motivational profiles and achievement in physical education. For both studies, we assessed achievement variables at different moments during the 10-week cycle. A longitudinal design is particularly appropriate because motivation is likely to have an impact on the variables of interest through time. Achievement was evaluated thanks to three objective criteria instead of the self-reported measures usually used in educational studies. First, we used a videotape to evaluate the effort exerted by the students during a task completion (Study 2). Second, we used the same procedure to assess students' initial performance in gymnastics (Study 2), as well as their performance 10 weeks later (Studies 1 and 2). Finally, the grade obtained for the class was provided by the teachers (Study 2).

A first hypothesis concerned the relationship between motivational profile at the beginning of the cycle and final performance in the activity at the end of the cycle. In Study 1, motivational profile was actually found to be significantly linked to students' final performance. The more that students displayed a self-determined profile of motivation, the higher was their performance. This positive link was previously observed in education (e.g., Miserandino, 1996), but with a cross-sectional design. The analyses conducted in Study 2 indicated that the students' motivational profile was significantly linked to their final performance, even after controlling for their initial performance. This is particularly important, because even if motivation was measured before performance in previous research, very little effort has been made to disentangle this effect from the effect of initial performance. Our result confirmed that no matter the initial level of skills of a student at the beginning of a teaching cycle, the more the student displayed a self-determined profile of motivation toward the activity, the higher would be his or her final performance.

We also examined the link between students' motivational profile and the effort they exerted. This link was significant when comparing the more self-determined students with the less self-determined ones. Students showing a moderate profile provided tangentially more effort than did students with a non-self-determined profile, but they were not distinguishable from students with a self-determined profile. Moderate levels of external regulation and amotivation combined with moderate levels of other motivations were thus found not to have too many negative consequences: It seems that this configuration is likely to promote effort in situations where self-consciousness is made salient, for example, when the performance is videotaped. A positive link was already observed in physical education between self-determined motivation and effort reported by the students themselves (e.g., Ntoumanis, 2001, 2002) or by their teacher (e.g., Standage et al., 2006). The pattern of result observed in our study confirmed this past literature but with a more objective measurement, because the effort of every student was evaluated with a videotape.

Finally, we examined the final grade obtained by the students. We expected that teachers' evaluations would first depend on the other indicators of achievement, namely the students' final perfor-

mance in the activity and the effort they exerted during the cycle. Indeed, teachers have been found to prefer more invested, hard-working students, whatever their level of performance (Covington & Omelich, 1979). This hypothesis was confirmed: Both indicators were significantly related to students' grades. It means that teachers gave better grades to students whom they perceived to work harder, even if they did not reach the greatest level of achievement. Motivational profile showed an indirect association with grade, thanks to its links with the effort provided and the final level of performance. Motivational profile was also found to have positive, direct links with the final grade given by the teacher, as well as the initial performance in the activity. This result suggests that teachers took into account some early characteristics of the students to evaluate them at the end of the cycle. One explanation of this result could be teachers' expectancies (e.g., Jussim & Eccles, 1992) and more precisely perceptual biases in their evaluation. This phenomenon is based on the influence teachers' early expectancies can have on their judgment of students' performance (e.g., Jussim & Eccles, 1995). Basically, if teachers have higher expectancies toward students who show a good performance at the beginning of the cycle, this could lead them to overevaluate their final level (see Trouilloud et al., 2002). Nevertheless, this hypothesis cannot be verified because teachers' variables were not measured in this study.

### *Implications for Self-Determination Theory and Research*

This work constitutes a substantial contribution to the SDT literature. Indeed, the representation of the multiple forms of motivation proposed by this theory represents one relatively unexamined issue to date. The motivational profiles observed in these studies and their links with students' achievement suggest interesting issues regarding the regulatory processes involved in education. The first issue to be addressed concerns the particularly negative pattern observed for students showing a non-self-determined profile. The association of high levels of external regulation and amotivation, associated with low level of every other form of motivation, was found to lead to poor effort, performance, and grade. A similar profile was previously observed by Ntoumanis and was associated with high levels of boredom and very low levels of enjoyment and effort (Ntoumanis, 2002). In the sport context, negative independent effects of the degree of external regulation and amotivation have been observed on short- and long-term persistence (Pelletier et al., 2001). The students displaying such motivational profiles are likely to "suffer from a lack of motivation and a sense of helplessness outside of the specific situations in which extrinsic rewards are available" (Lepper & Henderlong, 2000, p. 295). One might argue that, with the exception of when they are evaluated by their teacher, those students might have no motivation at all in class. They are likely to regularly adopt maladaptive strategies and avoid getting involved in the tasks proposed.

Our studies also underscore the particularly adaptive character of a motivational profile combining high scores of intrinsic motivation and identified regulation but also moderate levels of introjection and low levels of external regulation and amotivation. The importance of combining intrinsic and identified regulations was already underlined by Koestner and his colleagues (Burton et al., 2006; Koestner & Losier, 2002). They argued that possessing high levels of both types of regulation could allow one the flexi-



bility to adapt to a wide array of situations. More precisely, combining high levels of both intrinsic motivation and identified regulation is likely to lead to high levels of well-being and performance (Burton et al., 2006). More surprisingly, whereas it is considered a non-self-determined form of motivation, a moderate level of introjection was finally associated with the higher level of achievement. In spite of the negative affective states it implies, such as anxiety or guilt (Blais, Sabourin, Boucher, & Vallerand, 1990), this kind of regulation could likely enhance behavioral involvement. It was found to be related to persistence in previous work (Ntoumanis, 2002; Pelletier et al., 2001; Vallerand et al., 1997), but this relationship was observed only for a short-term period (e.g., Pelletier et al., 2001). The period of our studies was also relatively brief because of the 10-week teaching cycles. It would be interesting to investigate the educational outcomes of motivational profiles in the long term in order to verify if this motivational configuration is beneficial for achievement during one or several educational years.

It is important to note that a moderate level of introjection in itself does not guarantee those beneficial impacts on achievement. Indeed, students showing similar levels of introjection, associated with moderate levels of other kinds of regulation (i.e., what we labeled moderate motivational profile), did not show identical levels of performance in gymnastics, and they obtained lower grades. This result suggests that a moderate feeling of guilt or shame, as shown by the student motivated by introjection, does not necessarily have negative consequences, if at the same time the student gets a certain satisfaction from the activity (i.e., intrinsic motivation) and anticipates that the activity would help him or her to reach personal goals (i.e., identified regulation). Conversely, when these feelings are not associated with interest or personal value for the activity, the outcomes could be less positive.

Globally our results suggest an adaptive role of the self-determined motivational profile, as shown by students with high levels of intrinsic motivation toward stimulation, accomplishment, and knowledge; high levels of identified regulation; and also moderately high levels of introjection. These elements underscore the importance of the internalization process, thanks to which individuals do not act without feelings of control or competence and without reacting to external contingencies, that is, they do not show amotivation or external regulation. This result is coherent with the propositions of Koestner and Losier (2002), who assumed that it is appropriate to develop both intrinsic and identified goals. This "dual motivational system" would promote the pursuit of both short-term goals, thanks to "the energizing emotions such as interest and excitement" that it implies, and the pursuit of long-term goals that identification is more likely to enhance, because it gives "significance [to] one's current pursuits and fosters positive emotions such as pride in one's accomplishments in the domain" (p. 115).

Past research conducted in education has examined how the social context and more particularly how teachers could influence the degree of satisfaction of students' basic needs and motivations, depending on the motivational climate they generate. The teachers who support autonomy are more likely to generate the internalization of the activity among students. They spend more time listening to their students and acknowledge their perspective, they provide more support for the quality of students' performance and progress, and they promote more choice and initiative and participation in decisions (Grolnick & Ryan, 1989). In contrast, teachers

who are controlling use more directives; give more solutions to the students; criticize them more; and put more pressure on them using rewards, threats, and deadlines. There is empirical evidence showing that the degree to which teachers are autonomy supportive versus controlling is significantly linked with students' need satisfaction and motivations (see Reeve, 2002, for a review), especially in physical education (e.g., Ntoumanis, 2001; Standage et al., 2003).

Finally, Koestner and Losier (2002) also recommend providing a structure to promote self-determined forms of motivation where students have at their disposal consistent guidelines, rules, and expectations relative to their behavior and a rationale for the tasks and activities proposed. These authors also advanced that both autonomy support and structure are needed to lead to the internalization of the activity. Even if the educational climate was not taken into account in the present study, past literature has suggested that a teacher showing characteristics of an autonomy supportive climate and providing structure to the students would contribute to the development of a self-determined motivational profile beneficial for achievement.

### *Limitations and Research Perspectives*

These studies are not exempted from a certain number of limitations. First, the validity of the effort measure can be questioned, since the camcorder was visible. It is thus possible that a social desirability phenomenon occurred, leading some students to put more effort in the task than they usually did. Perhaps this can account for the low percentage of variance of this variable explained by the motivational profile, as well as the absence of significance for the path between effort and final performance. Second, the length of the cycle was relatively short, and the effects of motivation observed in 10-week cycles would probably be more important in studies carried out during a whole school year or with a follow-up of students over several years. Indeed, if students practice the same activities each year, the beneficial or detrimental effect of motivational profiles on achievement could be cumulative and lead to greater disparities between students at the end of high school, even if they had the same performance at the beginning. Moreover, the generalization of these results can be questioned, and further studies could replicate those findings in other activities or educational disciplines. Finally, as in many naturalistic studies, we cannot exclude that some of the relationships observed are due to the omission of a relevant variable (see, e.g., Judd & McClelland, 1989). For instance, the educational climate provided by the teacher might have had an impact on students' motivation and achievement at the same time. However, given that students' motivation was assessed during the first lesson of the cycle, the potential effect of the teacher climate is limited. Further studies should nevertheless control this variable.

### *References*

- Amabile, T. M. (1985). Motivation and creativity: Effects of motivational orientation on creative writers. *Journal of Personality and Social Psychology, 18*, 393-397.
- Amabile, T. M., Hill, K. G., Hennessey, B. A., & Tighe, E. M. (1994). The Work Preference Inventory: Assessing intrinsic and extrinsic motivational orientations. *Journal of Personality and Social Psychology, 66*, 950-967.
- Biddle, S. J. H., & Goudas, M. (1997). Effort is virtuous: Teacher prefer-



- ences of pupil effort, ability and grading in physical education. *Educational Research*, 39, 350–355.
- Biddle, S. J. H., & Wang, J. C. K. (2003). Motivation and self-perception profiles and links with physical activity in adolescent girls. *Journal of Adolescence*, 26, 687–701.
- Blais, M. R., Sabourin, S., Boucher, C., & Vallerand, R. J. (1990). Toward a motivational model of couple happiness. *Journal of Personality and Social Psychology*, 59, 1021–1031.
- Boggiano, A. N., Main, D. S., & Katz, P. A. (1988). Children's preference for challenge: The role of perceived competence and control. *Journal of Personality and Social Psychology*, 54, 134–141.
- Braten, I., & Olaussen, B. S. (2005). Profiling individual differences in student motivation: A longitudinal cluster-analytic study in different academic contexts. *Contemporary Educational Psychology*, 30, 359–396.
- Brière, N. M., Vallerand, R. J., Blais, M. R., & Pelletier, L. G. (1995). Development and validation of a measure of intrinsic, extrinsic and amotivation in the sport context: The Échelle de Motivation dans les Sports (ÉMS). *International Journal of Sport Psychology*, 26, 465–489.
- Burton, K. D., Lydon, J. E., D'Alessandro, D. U., & Koestner, R. (2006). The differential effects of intrinsic and identified motivation on well-being and performance: Prospective, experimental, and implicit approaches to self-determination theory. *Journal of Personality and Social Psychology*, 91, 750–762.
- Chanal, J. P., Marsh, H. W., Sarrazin, P. G., & Bois, J. E. (2005). Big-fish-little-pond effects on gymnastics self-concept: Social comparison processes in a physical setting. *Journal of Sport and Exercise Psychology*, 27, 53–70.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Cokley, K. O. (2000). Examining the validity of the Academic Motivation Scale by comparing scale construction to self-determination theory. *Psychological Reports*, 86, 560–564.
- Cokley, K. O., Bernard, N., Cunningham, D., & Motoike, J. (2001). A psychometric investigation of the Academic Motivation Scale using a United States sample. *Measurement and Evaluation in Counseling and Development*, 34, 109–119.
- Covington, M. V., & Mueller, K. J. (2001). Intrinsic versus extrinsic motivation: An approach/avoidance reformulation. *Educational Psychology Review*, 13, 157–176.
- Covington, M. V., & Omelich, C. L. (1979). Effort: The double edged sword in school achievement. *Journal of Educational Psychology*, 71, 169–182.
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.
- Deci, E. L., & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Human needs and the self-determination theory. *Psychological Inquiry*, 11, 227–268.
- Deci, E. L., Vallerand, R. J., Pelletier, L. G., & Ryan, R. M. (1991). Motivation in education: The self-determination perspective. *Educational Psychologist*, 26, 325–346.
- Fairchild, A. J., Horst, S. J., Finney, S. J., & Barron, K. E. (2005). Evaluating existing and new validity evidence for the Academic Motivation Scale. *Contemporary Educational Psychology*, 30, 331–358.
- Ferrer-Caja, E., & Weiss, M. R. (2000). Predictors of intrinsic motivation among adolescent students in physical education. *Research Quarterly for Exercise and Sport*, 71, 267–279.
- Fortier, M. S., & Grenier, M. N. (1999). Déterminants personnels et situationnels de l'adhérence à l'exercice: Une étude prospective [Personal and situational determinants of exercise adherence: A prospective study]. *STAPS*, 48, 25–37.
- Good, T. L., & Brophy, J. E. (2000). *Looking into classrooms* (5th ed.). New York: Longman.
- Goudas, M., Biddle, S., & Underwood, M. (1995). A prospective study of the relationships between motivational orientations and perceived competence with intrinsic motivation and achievement in a teacher education course. *Educational Psychology*, 15, 89–96.
- Grolnick, W. S., & Ryan, R. M. (1989). Parent styles associated with children's self-regulation and competence in school. *Journal of Educational Psychology*, 81, 143–154.
- Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
- Härdle, W., & Simar, L. (2003). *Applied Multivariate Statistical Analysis*. Berlin, Germany: Springer Verlag.
- Hodge, K., & Petlichkoff, L. (2000). Goal profiles in sport motivation: A cluster analysis. *Journal of Sport and Exercise Psychology*, 22, 256–272.
- Jöreskog, K. G., & Sörbom, J. (2003). *Lisrel 8.54: User's reference guide*. Chicago: SSI.
- Judd, C. M., & McClelland, C. H. (1989). *Data analysis: A model-comparison approach*. Orlando, FL: Harcourt.
- Jussim, L., & Eccles, J. (1992). Teacher expectations II: Construction and reflection of student achievement. *Journal of Personality and Social Psychology*, 63, 947–961.
- Jussim, L., & Eccles, J. (1995). Are teacher expectations biased by students' gender, social class, or ethnicity? In Y. T. Lee, L. J. Jussim, & C. R. McCauley (Eds.), *Stereotype accuracy: Toward appreciating group differences* (pp. 245–271). Washington, D.C.: American Psychological Association.
- Koestner, R., & Losier, G. F. (2002). Distinguishing three ways of being highly motivated: A closer look at introjection, identification, and intrinsic motivation. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 101–121). Rochester, NY: University of Rochester Press.
- Lepper, M. R., & Henderlong, J. (2000). Turning "play" into "work" and "work" into "play": 25 years of research on intrinsic versus extrinsic motivation. In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 257–307). San Diego, CA: Academic Press.
- Levesque, C., Zuehlke, N., Stanek, L., & Ryan, R. (2004). Autonomy and competence in German and American university students: A comparative study based on self-determination theory. *Journal of Educational Psychology*, 96, 68–84.
- Li, F. (1999). The Exercise Motivation Scale: Its multifaceted structure and construct validity. *Journal of Applied Sport Psychology*, 11, 97–115.
- Marshall, S. J., Biddle, S. J. H., Sallis, J. F., McKenzie, T. L., & Conway, T. L. (2002). Clustering of sedentary behaviors and physical activity among youth: A cross-national study. *Pediatric Exercise Sciences*, 14, 401–417.
- McNeill, M. C., & Wang, C. K. (2005). Psychological profiles of elite school sports players in Singapore. *Psychology of Sport and Exercise*, 6, 117–128.
- Meece, J. L., & Holt, K. (1993). A pattern analysis of students' achievement goals. *Journal of Educational Psychology*, 85, 582–590.
- Miserandino, M. (1996). Children who do well at school: Individual differences in perceived competence and autonomy in above-average children. *Journal of Educational Psychology*, 88, 203–214.
- Mosteller, F., & Tukey, J. W. (1977). *Data analyses and regression: A second course in statistics*. Reading, MA: Addison-Wesley.
- Ntoumanis, N. (2001). A self-determination approach to the understanding of motivation in physical education. *British Journal of Educational Psychology*, 71, 225–242.
- Ntoumanis, N. (2002). Motivational clusters in a sample of British physical education classes. *Psychology of Sport and Exercise*, 3, 177–194.
- Otis, N., Grouzet, F. M. E., & Pelletier, L. G. (2005). The latent motivational change in academic setting: A three-year longitudinal study. *Journal of Educational Psychology*, 97, 170–183.
- Pelletier, L., Fortier, M., Vallerand, R., & Brière, N. (2001). Associations

- among perceived autonomy support, forms of self-regulation, and persistence: A prospective study. *Motivation and Emotion*, 25, 279–306.
- Pelletier, L. G., Fortier, M. S., Vallerand, R. J., Tuson, K. M., Brière, N. M., & Blais, M. R. (1995). Toward a new measure of intrinsic motivation, extrinsic motivation, and amotivation in sports: The Sports Motivation Scale (SMS). *Journal of Sport and Exercise Psychology*, 17, 35–53.
- Ratelle, C. F., Guay, F., Larose, S., & Sénécal, C. (2004). Family correlates of trajectories of academic motivation during a school transition: A semi-parametric group-based approach. *Journal of Educational Psychology*, 96, 743–754.
- Reeve, J. (2002). Self-determination theory applied to educational settings. In E. Deci & R. Ryan (Eds.), *Handbook of self-determination research* (pp. 183–203). Rochester, NY: The University of Rochester Press.
- Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology*, 57, 749–761.
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development and well-being. *American Psychologist*, 55, 68–78.
- Sansone, C., & Harackiewicz, J. (2000). Controversies and new directions—is it déjà vu all over again? In C. Sansone & J. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 443–453). San Diego, CA: Academic Press.
- Sarrazin, P., Vallerand, R., Guillet, E., Pelletier, L., & Cury, F. (2002). Motivation and dropout in female handballers: A 21-month prospective study. *European Journal of Social Psychology*, 32, 395–418.
- Standage, M., Duda, J., & Ntoumanis, N. (2003). A model of contextual motivation in physical education: Using constructs from self-determination and achievement goal theories to predict physical activity intentions. *Journal of Educational Psychology*, 95, 97–110.
- Standage, M., Duda, J., & Ntoumanis, N. (2005). A test of self-determination theory in school physical education. *British Journal of Educational Psychology*, 75, 411–433.
- Standage, M., Duda, J., & Ntoumanis, N. (2006). Students' motivational processes and their relationship to teacher rating in school physical education: A self-determination theory approach. *Research Quarterly for Exercise and Sport*, 77, 100–110.
- Trouilloud, D., Sarrazin, P., Martinek, T., & Guillet, E. (2002). The influence of teacher expectations on student achievement in physical education classes: Pygmalion revisited. *European Journal of Social Psychology*, 32, 591–607.
- Vallerand, R. J. (1997). Toward a hierarchical model of intrinsic and extrinsic motivation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (pp. 271–360). New York: Academic Press.
- Vallerand, R. J., & Bissonette, R. (1992). Intrinsic, extrinsic and amotivational styles as predictors of behavior: A prospective study. *Journal of Personality*, 60, 599–620.
- Vallerand, R. J., Blais, M., Brière, N., & Pelletier, L. G. (1989). Construction et validation de l'échelle de motivation en éducation (EME) [Construction and validation of the Educational Motivation Scale (EMS)]. *Revue Canadienne des Sciences du Comportement*, 21, 323–349.
- Vallerand, R. J., & Fortier, M. N. (1998). Measures of intrinsic and extrinsic motivation in sport and physical activity: A review and critique. In J. L. Duda (Ed.), *Advances in Sport and Exercise Psychology Measurement* (pp. 81–101). Morgantown, WV: Fitness Information Technology.
- Vallerand, R. J., Fortier, M. N., & Guay, F. (1997). Self-determination and persistence in a real-life setting: Toward a motivational model of high school dropout. *Journal of Personality and Social Psychology*, 72, 1161–1176.
- Vallerand, R. J., Pelletier, L. G., Blais, M. R., Brière, N. M., Sénécal, C., & Vallières, E. F. (1993). On the assessment of intrinsic, extrinsic, and amotivation in education: Evidence on the concurrent and construct validity of the academic motivation scale. *Educational and Psychological Measurement*, 53, 159–172.
- Vansteenkiste, M., Simons, J., Lens, W., Sheldon, K., & Deci, E. (2004). Motivating learning, performance, and persistence: The synergistic effects of intrinsic goal contents and autonomy-supportive contexts. *Journal of Personality and Social Psychology*, 87, 246–260.
- Vlachopoulos, S. P., Karageorghis, C. I., & Terry, P. C. (2000). Motivation profiles in sport: A self-determination theory perspective. *Research Quarterly for Exercise and Sport*, 7, 387–397.
- Wang, J. C. K., & Biddle, S. J. H. (2001). Young people's motivation profiles in physical activity: A cluster analysis. *Journal of Sport and Exercise Psychology*, 23, 1–22.
- Wang, J. C. K., Chatzisarantis, N. L. D., Spray, C. M., & Biddle, S. J. H. (2002). Achievement goal profiles in school physical education: Differences in self-determination, sport ability beliefs, and physical activity. *British Journal of Educational Psychology*, 72, 433–445.
- Weiner, B. (1979). A theory of motivation for some classroom experience. *Journal of Educational Psychology*, 71, 3–25.

Received March 21, 2006

Revision received July 3, 2007

Accepted July 5, 2007 ■



# Teachers' Occupational Well-Being and Quality of Instruction: The Important Role of Self-Regulatory Patterns

Uta Klusmann, Mareike Kunter, Ulrich Trautwein, Oliver Lüdtke, and Jürgen Baumert  
Max Planck Institute for Human Development

Teachers' occupational well-being (level of emotional exhaustion and job satisfaction) and quality of instruction are two key aspects of research on teaching that have rarely been studied together. The role of occupational engagement and resilience as two important work-related self-regulatory dimensions that predict occupational well-being and teachers' instructional performance in the classroom was investigated. In Part 1 of the study, self-regulatory data from 1,789 German mathematics teachers were subjected to a latent profile analysis, yielding four self-regulatory types (healthy-ambitious, unambitious, excessively ambitious, and resigned) that differed significantly on emotional exhaustion and job satisfaction. In Part 2, the association between teachers' self-regulatory type and instructional performance was examined in a subsample of 318 teachers. Results showed that teachers' self-regulatory type predicted the quality of instruction in three of the four aspects of instructional performance examined. Moreover, teachers' self-regulatory type was systematically linked to differences in students' motivation. No association was found between teacher self-regulation and student achievement.

*Keywords:* teachers, instructional quality, emotional exhaustion, job satisfaction, latent profile analysis

High rates of early retirement among teachers and disappointing student outcomes in recent international educational assessments have fueled concerns that many teachers in Western industrialized countries are not succeeding fully in their work (Organisation for Economic Cooperation and Development [OECD], 2005; Vandenberghe & Huberman, 1999).

But what are "successful" (or "effective") teachers? What are the outcomes that successful teachers are expected to produce? Research on successful teachers can be classified into two main strands that rarely overlap. The first, research on learning and instruction, focuses on the quality of instruction as a major outcome variable that in turn predicts students' achievement gains and motivational and personality development (Bromme, 2001). The major emphasis is on the teacher's role as an instructor, and successful teachers are identified as those who provide good instruction. The second strand, research on occupational health,

focuses on teachers' health and affective well-being. In this research tradition, successful teachers are characterized as experiencing low stress, showing no symptoms of burnout, and reporting high job satisfaction (Kyriacou, 2001; Maslach & Leiter, 1999).

In the present article, we propose that these two core aspects cannot be separated and that research on successful teachers will profit considerably from taking both research traditions and sets of outcome measures into account. Specifically, we investigated whether differences in occupational well-being and instructional performance can be explained by teachers' self-regulatory patterns. Guided by a theory-based research model, we focused on the intraindividual interplay of two self-regulatory behaviors that are highly relevant in the occupational domain: work engagement and resilience (Hallberg & Schaufeli, 2006; Schaarschmidt, Kieschke, & Fischer, 1999). Based on the concept of balanced commitment (Hallsten, 1993) and conservation of resources theory (Hobfoll, 1989), we argue that a combination of high engagement in the teaching profession (work engagement) with the capacity to emotionally distance oneself from work and cope with failure (resilience) is associated with both high levels of occupational well-being (low levels of exhaustion, high job satisfaction) and better instructional performance, and in turn leads to favorable student outcomes.

There are two parts to our empirical study. In Part 1, we empirically identified different patterns of engagement and resilience by applying latent profile analysis to a sample of more than 1,700 German mathematics teachers. Moreover, we examined the relationship between these self-regulatory patterns and two measures of occupational well-being: emotional exhaustion and job satisfaction. In Part 2, drawing on a subsample of 318 teachers and their students, we explored differences in the instructional performance of teachers of the four different self-regulatory types and examined whether these differences are systematically associated with their students' achievement and motivational experience.

---

Uta Klusmann, Mareike Kunter, Ulrich Trautwein, Oliver Lüdtke, and Jürgen Baumert, Max Planck Institute for Human Development, Berlin, Germany.

The research reported in this article is based on data from the Professional Competence of Teachers, Cognitively Activating Instruction, and the Development of Students' Mathematical Literacy study (COACTIV) directed by Jürgen Baumert (Max Planck Institute for Human Development, Berlin, Germany), Werner Blum (Kassel University, Germany), and Michael Neubrand (Carl von Ossietzky University, Oldenburg, Germany). The project is funded by the German Research Foundation (DFG; BA 1461/2-2) as part of its priority Program on School Quality (BIQUA).

The authors thank Stefan Krauss and Martin Brunner for their feedback on the present research and Susannah Goss for editorial assistance.

Correspondence concerning this article should be sent to Uta Klusmann, Center for Educational Research, Max Planck Institute for Human Development, Lentzeallee 94, 14195, Berlin, Germany. E-mail: klusmann@mpib-berlin.mpg.de

## Successful Teachers: Feeling Well and Performing Effectively in the Classroom

Our theoretical framework combines elements of the transactional perspective from health psychology (Lazarus & Folkman, 1987) with the extended process-product model and the expertise approach from research on learning and instruction (Bromme, 2001; Sternberg & Horvath, 1995). Features of the working environment and teachers' personal characteristics are conceived to be antecedents of teachers' occupational well-being, teachers' instructional performance in the classroom, and student outcomes (Maslach, Schaufeli, & Leiter, 2001; Shuell, 1996; Weinert & Helmke, 1995). Student achievement and motivational development are thus hypothesized to be affected by their experiences in the classroom, which are shaped by teachers' instructional performance. The relations proposed and variables considered are illustrated in Figure 1.

We do not examine the entire model in this article but focus instead on the relationship between selected teacher characteristics (i.e., profiles of self-regulatory behavior) and two aspects of success (i.e., teachers' occupational well-being and instructional quality). In the following sections, we first describe our conceptualizations of occupational well-being and instructional performance as criteria for successful teachers and then turn to the patterns of self-regulation that we consider crucial for obtaining these outcomes.

### Aspects of Teachers' Occupational Well-Being

Occupational well-being can be defined as the "positive evaluation of various aspects of one's job, including affective, motivational, behavioral, cognitive and psychosomatic dimensions" (Van Horn, Taris, Schaufeli, & Schreurs, 2004, p. 366). To date, research on teachers has tended to focus on the negative side of

well-being by exploring concepts such as stress, burnout, and low job satisfaction (Maslach et al., 2001; Schaufeli & Enzmann, 1998).

Kyriacou (2001, p. 28) defined teacher *stress* "as the experience by a teacher of unpleasant, negative emotions, such as anger, anxiety, tension, frustration or depression, resulting from some aspect of their work as a teacher." The experience of chronic stress is one of the core dimensions of burnout, a diagnosis that gained currency in the context of work-related strain in social workers. The classic and most popular definition of burnout was proposed by Maslach (Maslach, Jackson, & Leiter, 1996), who described it as a psychological syndrome characterized by three symptoms: emotional exhaustion, depersonalization, and reduced personal accomplishment. *Emotional exhaustion*, the most obvious manifestation and central quality of burnout, involves feelings of being emotionally drained and depleted of emotional resources. *Depersonalization* is characterized by a negative, callous, and detached attitude to others (in this case, students). *Reduced personal accomplishment* implies feelings of incompetence and a negative self-evaluation of job performance. In line with Rudow (1999) and Schaufeli and Enzman (1998), we regard burnout as a result of prolonged job stress. In the present study, we assessed the emotional exhaustion that teachers report with respect to their work, thus tapping a key dimension of burnout and a major operationalization of chronic stress.

In occupational research, the concept of *job satisfaction* is one of the most studied aspects of job-related well-being, reflecting "a positive (or negative) evaluative judgment one makes about one's job or job situation" (Weiss, 2002, p. 175), which can be explained by the perceived degree of need fulfillment within the organizational setting. The nature of the relationship between job satisfaction and stress is not yet clear: Whereas some researchers regard burnout as a cause for low job satisfaction, others assume the

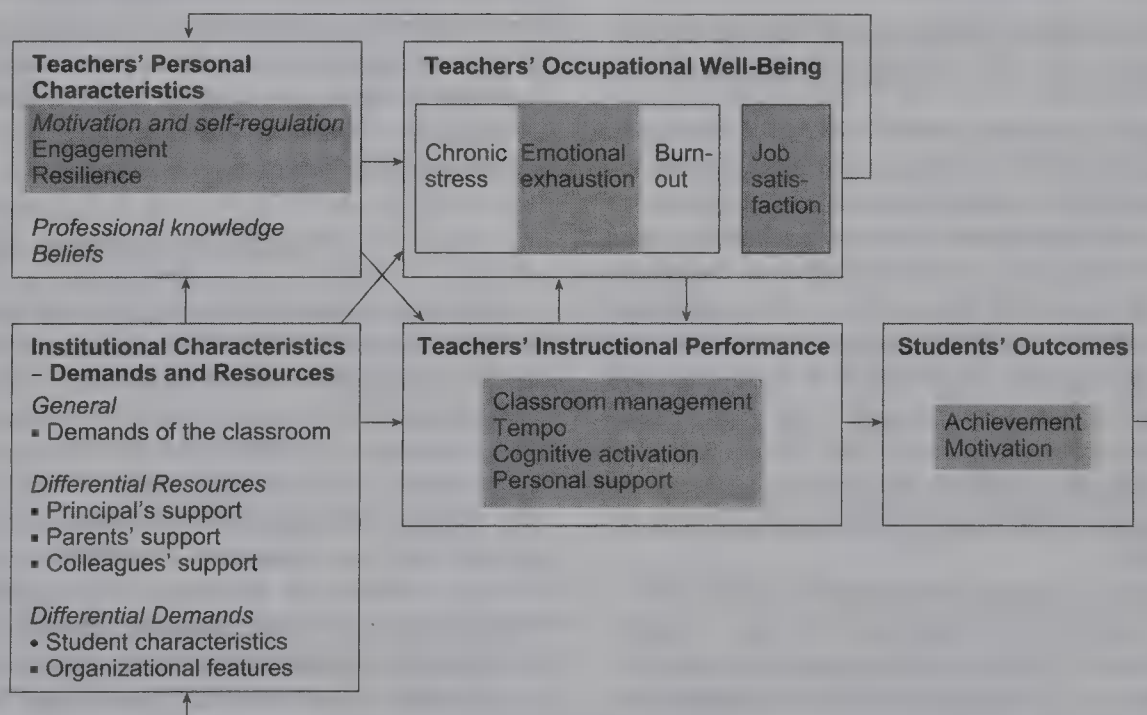


Figure 1. Heuristic working model of teacher characteristics, teachers' occupational well-being, and instructional performance. Only the shaded sections were examined.



opposite mechanism. Empirically, there is some evidence that low job satisfaction is a consequence of perceived stress at the workplace (Wolpin, Burke, & Greenglass, 1991).

Stress and job satisfaction reflect different aspects of teachers' occupational well-being (Van Horn et al., 2004). What they have in common is that their undesirable forms (i.e., high levels of stress and exhaustion and low levels of satisfaction) have negative consequences for individual psychological and physical health (Melamed, Shirom, Toker, Berliner, & Shapira, 2006) and for the organization itself in terms of lower performance and higher fluctuation (Judge, Thoresen, Bono, & Patton, 2001).

### *Teachers' Instructional Performance in the Classroom*

Teachers' main field of activity is the classroom. The crucial question is whether teachers succeed in creating challenging and adaptive learning situations and in effectively guiding their students through the learning process (Collins, Greeno, & Resnick, 2001)—in other words, whether their instructional behavior is successful.

Successful teachers can be described as performing well on at least four aspects of instructional behavior (see Kunter et al., 2007; Shuell, 1996). First, they succeed in establishing a well-structured environment characterized by low disturbance levels and efficient use of time (classroom management; Emmer & Stough, 2001). Second, they foster cognitive activation, enabling students to develop new insights and understandings (Turner et al., 1998). Third, successful teachers proceed at an appropriate pace that neither overtaxes nor underestimates students. Fourth, successful teachers are able to create a supportive social environment (e.g., show patience with students' mistakes, take time to talk about interpersonal problems, and behave fairly in critical situations) in which students receive personal guidance and feel personally valued (e.g., Ryan & Powelson, 1991). All four aspects of instructional quality are thought to be related to students' learning outcomes.

### *Personal Characteristics Contributing to Success in the Teaching Profession: Engagement and Resilience*

According to our theoretical framework, successful teachers are those who experience high levels of occupational well-being and succeed in creating optimal learning environments for their students. Several teacher characteristics have been proposed as being conducive to these outcomes. Research on health-related outcomes has focused on broad personality characteristics, such as emotional stability, internal locus of control, hardiness, self-efficacy, and coping styles (for an overview, see Schaufeli & Enzmann, 1998; Semmer, 1996). At the same time, research on instructional performance and student achievement has focused on cognitive characteristics, such as teachers' professional knowledge, beliefs, and self-efficacy (Shulman, 1986; Staub & Stern, 2002; Tschannen-Moran & Hoy, 2001).

We hypothesized two work-related personal characteristics to foster an effective approach to occupational demands: work engagement and resilience. Previous studies have shown both constructs to be important in teaching as well as in other professions. Work engagement is the willingness to invest energy and resources in one's job and is described as a "positive, fulfilling work-related state of mind" (Hallberg & Schaufeli, 2006), with high levels of

vigor (energy, effort) and dedication (in terms of inspiration, pride, and a sense of significance). High work engagement is associated with the willingness to stay in a particular organization and with high performance levels (Hakanen, Bakker, & Schaufeli, 2006). Resilience reflects the individual's reaction to work-related demands and describes the ability to deal with failure and to maintain a healthy distance with one's work (Schaarschmidt et al., 1999). It includes emotional distancing, a low tendency to give up after failure, active coping, and mental stability. Kyriacou (2001) described it as an active and palliative coping style.

Research on occupational behavior suggests that the most adaptive self-regulatory pattern is characterized by high levels of both engagement and resilience. In his theory of balanced commitment, Hallsten (1993) described individuals who show low vulnerability to stress as being able to distance themselves from their work and, at the same time, are highly engaged and committed. Similarly, conservation of resources theory (Hobfoll, 1989) states that well-being arises from a balance between the processes of investing resources (engagement) and recovering resources (resilience). Empirical evidence in support of this claim was reported by Schaar-schmidt et al. (1999), who used a cluster-analytic procedure to identify four different patterns of self-regulation in teachers. The healthy-ambitious (H) type, with high scores on both occupational engagement and resilience, is seen as the best adapted pattern. The unambitious (U) type is characterized by low occupational engagement but high resilience. The remaining two types are thought to be at high risk for burnout and stress. The excessively ambitious (A) type, scoring high on engagement and low on resilience, is characterized by excessive engagement, striving for perfection, and an inability to recover emotionally from work. The resigned (R) type is characterized by low engagement and low stress resistance.

Schaarschmidt et al. (1999) found that individuals of the H type reported less physical and psychological strain, had lower absence rates, and had lower means on the three burnout symptoms than the two at-risk types, with at-risk R type faring worse than at-risk A type. Moreover, teachers of the four self-regulatory types differed in their self-reported occupational knowledge and performance. Teachers of the H type ascribed to themselves more content and pedagogical knowledge than did teachers of the other self-regulatory types.

### *The Present Investigation*

In the present investigation, we focused on the intraindividual interplay of work engagement and resilience as teacher characteristics hypothesized to affect teachers' occupational well-being and instructional performance. Based on earlier studies, and in line with conservation of resources theory and the concept of balanced commitment (Hallsten, 1993; Hobfoll, 1989), we expected to find four different self-regulatory types. Moreover, we expected teachers who are both engaged and resilient to score highest on the outcome measures of well-being and instructional performance. We assumed that these teachers invest resources in their work and find meaning in their profession, but are at the same time able to distance themselves emotionally and to cope with failure. Beyond this H type, we expected the other three types to show distinct patterns of results on the outcome measures. In particular, engagement was expected to be more strongly associated with instruc-



tional performance than with well-being. In contrast, teachers scoring high on resilience (i.e., who are better able to distance themselves from work and to adopt an active coping style that protects them against exogenous stressors) were expected to score higher on measures of well-being.

We tested our assumptions in two sets of analyses. First, we empirically explored the four different patterns of self-regulation and related them to occupational well-being. Second, we examined how teachers' engagement and resilience are related to instructional quality and to students' motivational and achievement outcomes.

### Part 1: Four Profiles of Self-Regulatory Behavior and Their Association With Occupational Well-Being

In Part 1 of our study, we first investigated whether it is possible to empirically identify the four patterns of self-regulatory behavior described earlier (H, U, A, and R). To this end, we used latent profile analysis to categorize individuals into subgroups on the basis of the patterns of associations among variables indicating engagement and resilience. The model-based statistical approach of latent profile analysis compares favorably with the more widely used cluster-analytical approach in that it provides indices of the fit of any model tested, making it possible to compare models with different numbers of clusters (Vermunt & Magidson, 2002).

Second, we investigated whether these four teacher types differed in terms of emotional exhaustion and job satisfaction. Our four hypotheses were as follows: First, we expected teachers scoring high on both resilience and engagement (H type) to report the highest levels of job satisfaction and lowest levels of stress. Second, we expected teachers scoring high on resilience and low on engagement (U type) to report lower job satisfaction and more emotional exhaustion than H-type teachers. Although they are highly distanced from their work, these teachers may find no real meaning in their job. At the same time, given their high resilience scores, U-type teachers were expected to be more satisfied and less emotionally exhausted than the A-type and the R-type teachers. Third, we expected teachers of the A type, who are characterized by high engagement and low resilience, to show low job satisfaction and high emotional exhaustion. According to the principles of Hobfoll's (1989) conservation of resources theory, these teachers can be seen as trapped in a "loss spiral," investing a great deal of energy without experiencing sufficient gratification. Fourth, teachers with a profile of low engagement and resilience (R type) were expected to score lowest on job satisfaction and highest on emotional exhaustion. These individuals seem to have few resources to help them deal with work-related demands, resulting in exhaustion and dissatisfaction.

### Method

#### Sample

Data for the present research were provided by the Professional Competence of Teachers, Cognitively Activating Instruction, and the Development of Students' Mathematical Literacy study (COACTIV), which was embedded in the German extension to the 2003 cycle of the OECD's Programme for International Student Assessment (PISA). Within the German 2003 PISA assessment,

197 schools were drawn to form a representative sample of ninth-grade students. COACTIV surveyed up to 12 mathematics and science teachers in each of these schools, with teacher questionnaires being administered to assess biographical data, self-regulation, and occupational well-being. We received 1,789 completed questionnaires from the 1,940 teachers approached. On average, 9 teachers per school participated (range: 2–12 teachers). These teachers (47.8% male; age range: 25–65 years,  $M = 47.3$ ,  $SD = 9.4$ ) formed the basis for Part 1 of our study. The length of teaching experience ranged from 1 to 44 years ( $M = 20.6$ ,  $SD = 10.6$ ). All of the German secondary school tracks were represented.<sup>1</sup>

### Measures

**Engagement and resilience.** The work engagement and resilience dimensions were measured using eight scales from the Occupational Stress and Coping Inventory (AVEM; Schaar-schmidt et al., 1999). Means, standard deviations, and internal consistencies of the scale scores are provided in Table 1. Prompted by the instruction "We would like you to describe some of your typical behaviors, attitudes, and habits with respect to your working life," teachers were asked to rate their agreement with each item (four items per scale) on a 5-point response scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *partly agree*, 4 = *agree*, 5 = *strongly agree*). *Work engagement* was tapped by the following subscales: subjective significance of work, career ambitions, exertion, and perfectionism. *Resilience* was assessed by the subscales emotional distancing, low tendency to give up after failure, active coping, and mental stability. As shown in Table 1, all scale scores demonstrated good to high internal consistency ( $\alpha$  ranged from .79 to .82).

**Emotional exhaustion.** Emotional exhaustion was measured by an established German adaptation (Enzmann & Kleiber, 1989) of the Maslach Burnout Inventory (Maslach et al., 1996). Participants were asked to rate their agreement with five statements (e.g., "I often feel exhausted at school") on a 4-point response scale (1 = *strongly disagree*, 4 = *strongly agree*). Internal consistency of the scale scores was good ( $\alpha = .80$ ).

**Job satisfaction.** We used a short German version (Merz, 1979) of the Work Satisfaction Scale of the Job Diagnostic Survey (Hackman & Oldham, 1975) to assess job satisfaction. This measure focuses on global satisfaction with work rather than on certain facets of the job (e.g., "Given the choice, I would definitely become a teacher again"). Teachers rated six items on a 4-point response scale (1 = *strongly disagree*, 4 = *strongly agree*). The scale score showed high internal consistency ( $\alpha = .84$ ).

### Statistical Analysis

Latent profile analysis is the procedure best suited to our aim of identifying qualitatively distinct self-regulatory patterns (Lubke &

<sup>1</sup> The German secondary school system comprises several tracks that differ in terms of student achievement levels, years of schooling, and the academic opportunities available after graduation. A broad distinction can be drawn between the academic track (Gymnasium) and the nonacademic tracks (e.g., Realschule, Hauptschule), in that the final Gymnasium qualification (the *Abitur*) is obligatory for university entrance. In the present sample, 30.4% of the schools belong to the academic track.



Table 1  
*Descriptive Results for the Self-Regulatory, Emotional Exhaustion, and Job Satisfaction Scales*

Scale	N <sub>1</sub>			N <sub>2</sub>			N items
	M	SD	α	M	SD	α	
Work engagement (scale range: 1–5)							
Significance of work	2.57	0.79	.82	2.55	0.78	.81	4
Career ambitions	2.79	0.78	.81	2.80	0.80	.82	4
Exertion	2.96	0.79	.79	2.96	0.80	.82	4
Perfectionism	3.45	0.72	.79	3.42	0.65	.73	4
Resilience (scale range: 1–5)							
Emotional distancing	2.95	0.80	.82	2.98	0.83	.84	4
Low tendency to give up <sup>a</sup>	3.47	0.70	.80	3.47	0.67	.81	4
Active coping	3.39	0.60	.81	3.39	0.59	.82	4
Mental stability	3.34	0.74	.79	3.35	0.77	.80	4
Emotional exhaustion (scale range: 1–4)	2.10	0.62	.84	2.13	0.62	.82	5
Job satisfaction (scale range: 1–4)	2.95	0.71	.84	2.92	0.71	.84	6

Note. N<sub>1</sub> = 1,789; N<sub>2</sub> = 318.

Muthén, 2005; see Pastor, Barron, Miller, & Davis, 2007, for a didactical application). As a particular kind of mixture model, latent profile analysis can be applied to data that are assumed to consist of unobservable subgroups of individuals with different probability distributions. Although cluster analysis techniques are commonly used to divide persons into homogeneous subgroups, latent profile analysis offers several advantages over the traditional approach (see Pastor et al., 2007). Whereas cluster analysis is an exploratory technique, latent profile analysis is a model-based procedure that allows for more flexible model specification. In fact, it has been suggested that traditional cluster analysis is equivalent to a very restricted specification of a latent profile model (Vermunt & Magidson, 2002). Moreover, the fit indices provided in latent profile analysis enable different models to be compared and informed decisions to be made regarding the number of underlying classes.

The key difference between latent profile analysis and the common factor model is that “the common factor model decomposes the covariances to highlight relationships among the variables, whereas the latent profile model decomposes the covariances to highlight relationships among individuals” (Bauer & Curran, 2004, p. 6). In other words, factor models cluster items, whereas latent profile analyses cluster individuals (Lubke & Muthén, 2005). Because we were interested in patterns of self-regulatory behavior among teachers, we chose latent profile analysis to identify latent clusters of teachers.

We used three criteria to identify the number of latent clusters. First, we applied the Lo–Mendell–Rubin likelihood ratio test (LMR test; Lo, Mendell, & Rubin, 2001). This test compares the improvement in fit between neighboring class models (i.e., comparing the *k*-1 and *k*-cluster models) and provides a *p* value that can be used to determine whether there is statistically significant improvement in fit when an additional class is included (for applications, see Lubke & Muthén, 2005). Second, we evaluated the different cluster solutions in terms of their interpretability (mean patterns of clusters, differentiation of profiles, sizes of clusters). Third, we considered three commonly used information criterion indices based on the log likelihood of a fitted model and a penalty term for the number of model parameters and/or sample

size. Akaike’s information criterion (AIC) and Bayes’s information criterion (BIC) balance the improvement in fit associated with adding classes to the model against the number of parameters in the model. A decrease in these indices when an additional cluster is added indicates an improvement in model fit. Lubke and Muthén (2005) have proposed an adjusted BIC (aBIC), which applies a different penalty term in the equation for sample size.

All latent profile analyses reported in the present article were conducted using Mplus, Version 3 (Muthén & Muthén, 1998–2004).

Results

Identifying Different Self-Regulatory Styles

Latent profile analysis was conducted using the means of the eight scales assessing occupational engagement and resilience. Five models were specified, starting with a one-class model that estimated 16 parameters (8 variances and 8 intercepts). For each additional latent class, 17 new parameters were estimated (8 variances, 8 intercepts, and 1 class probability parameter). Table 2 presents the fit indices for the five latent profile models. Whereas the information criteria (AIC, BIC, and aBIC) decreased when additional latent classes were added, the LMR test provided clear support for the expected four-class solution. Because Lubke and Muthén (2005) suggested that content-oriented criteria be taken into account when specifying the number of latent classes, we inspected the mean patterns of the different solutions. The *z* scores of the eight scales assessing engagement and resilience (Figure 2) supported a four-class solution.

Class 1, which corresponded to the pattern of the H type (29.2% of participating teachers), was characterized by high scores on both the engagement and the resilience subscales (relative to the mean of the total sample). Class 2 conformed to the expected profile of the U type (25.4%) and was characterized by very low scores on the engagement scales and high scores on the four resilience scales. The third class, representing the A type (16.4%), was characterized by very high scores (more than one standard deviation above the overall group mean) on the engagement scales and

Table 2

*Fit Indices for Different Class Solutions: Latent Profile Analysis for the Class-Dependent Variance Model*

Model	No. of parameters	AIC	BIC	Sample-adjusted BIC	<i>p</i> LMR
1-class	16	32176.67	32264.50	32213.67	—
2-class	33	30338.93	30520.08	30415.24	.00
3-class	50	29666.28	29940.75	29781.90	.02
4-class	67	29184.12	29551.91	29339.05	.03
5-class	84	28911.41	29372.53	29105.66	.72

*Note.*  $N = 1,789$ . The 4-class solution was chosen for the subsequent analyses on the basis of the fit indices and profiles of the means. AIC = Akaike's information criterion; BIC = Bayes's information criterion; *p* LMR = Lo-Mendell-Rubin likelihood ratio test for  $n$  versus  $n - 1$  classes.

low resilience scores. Finally, the pattern of results for the fourth class conformed to the profile expected for the R type (29.0%), with low scores on both engagement and resilience.

We next explored whether there were any differences in gender, school track, age, and years of teaching experience among individuals assigned to the different types. Chi-square tests revealed differences in the gender distribution of the four types,  $\chi^2(3, N = 1,740) = 45.74, p < .05$ . In terms of effect sizes (Grissom & Kim, 2005), the gender differences among the four patterns can be seen as small ( $r = .16$ ), with a higher frequency of women in the R (34.9% of women vs. 22.8% of men) and A (18.3% of women vs. 14.7% of men) types, and a higher frequency of men in the U (29.7% of men vs. 20.9% of women) and H (32.7% of men vs. 25.9% of women) types. Further, we found differences in the frequency of academic- and non-academic-track teachers assigned to the four self-regulatory patterns,  $\chi^2(3, N = 1,779) = 11.44, p < .05$ . In terms of effect sizes, these differences can be seen as rather small ( $r = .08$ ), with a higher frequency of academic-track teachers in the A type (20.6% of academic-track teachers vs. 14.5% of

non-academic-track teachers) and a higher frequency of non-academic-track teachers in the R type (30.3% of non-academic-track teachers vs. 26.0% of academic-track teachers). Analysis of variance revealed a statistically significant but rather small overall effect for age,  $F(3, 1693) = 2.82, p < .05, \eta^2 = .005$ . A post hoc analysis (Student-Newman-Keuls) did not find statistically significant differences between any of the four groups, however. Similarly, there was no overall effect for years of teaching experience,  $F(3, 1698) = 1.96, p > .05, \eta^2 = .003$ .

#### *Self-Regulatory Styles and Their Association With Emotional Exhaustion and Job Satisfaction*

Did individuals of the four self-regulatory types differ in terms of the level of occupational well-being they reported? The correlational associations between the self-regulatory scales and the two well-being measures are reported above the diagonal in Table 3.

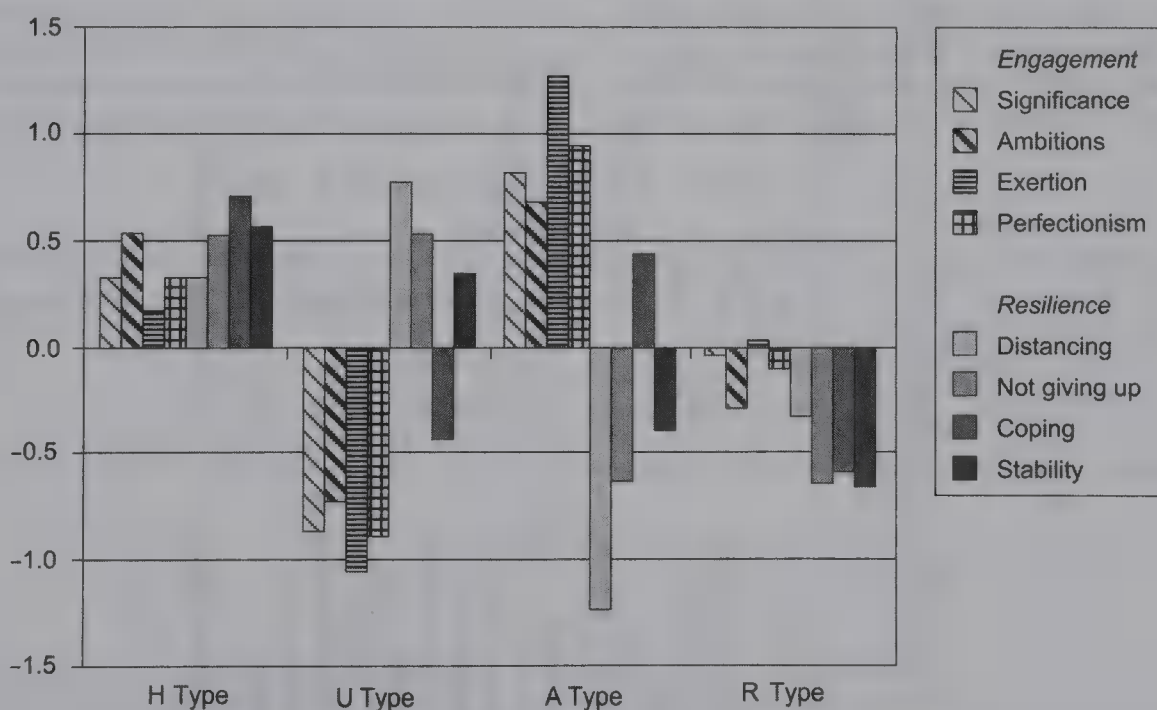


Figure 2. Z scores for subscales of engagement and resilience by self-regulatory type. H = healthy-ambitious type; U = unambitious type; A = excessively ambitious type; R = resigned type.



Table 3  
*Intercorrelations Between Self-Regulatory Dimensions (1–8), Teachers' Occupational Well-Being (9–10), Instructional Performance (11–14), and Student Outcomes (15–16)*

Scale	Self-regulatory dimensions								Occupational well-being		Instructional performance				Student outcomes	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 Significance of work																
2 Career ambitions	.50**															
3 Exertion	.45**	.35**														
4 Perfectionism	.28**	.35**	.44**													
5 Emotional distancing	-.36**	-.25**	-.54**	-.31**												
6 Low tendency to give up	-.13	.01	-.23**	-.15**	.45**											
7 Active coping	.14	.38**	.19**	.34**	.12	.33**										
8 Mental stability	-.03	.14	-.09	.05	.30**	.51**	.31**									
9 Emotional exhaustion	.01	-.13	.20**	.02	-.41**	-.44**	-.26**	-.37**								
10 Job satisfaction	.10	.11	-.12	-.09	.32**	.35**	.18**	.26**	-.56**							
11 Classroom management	-.01	-.06	.05	.00	.07	-.04	.17	.04	-.19**	.15**						
12 High tempo	-.11	-.18**	-.10	-.03	.00	-.03	-.08	-.21**	.13**	-.16**	-.26**					
13 Cognitive activation	.03	.09	.12**	-.03	.07	.09	.10	.07	-.09	.12**	.36**	-.36**				
14 Personal support	.12**	.20**	.13**	.03	.04	.05	.13**	.20**	-.10**	.17**	.20**	-.70**	.68**			
15 Mathematical achievement	-.02	.05	.08	-.05	-.04	.03	-.07	-.04	-.07	.06	.25**	-.21**	-.04	-.13**		
16 Student basic need satisfaction	.07	.08**	.04	-.07	.04	.16**	.07	.16**	-.10	.18**	.17**	-.42**	.61**	.67**	-.33**	

Note. Above the diagonal:  $N = 1,789$ ; below the diagonal:  $N = 318$ .

\*\*  $p < .01$ .

We performed analyses of variance (ANOVA) with post hoc tests (Student–Newman–Keuls) and analyses of variance and covariance (ANCOVA) to compare mean emotional exhaustion and job satisfaction across the four groups. A Bonferroni-type adjustment was made to avoid inflated Type I error through multiple testing. With a familywise error rate ( $\alpha$ ) of .05 and four significance tests (two ANOVAs and two ANCOVAs), the critical value for each test ( $\alpha$ ) was adjusted to .0125. The ANOVA (see Table 4) revealed substantial differences in the emotional exhaustion reported by the four groups,  $F(3, 1785) = 116.51, p < .001$ . At  $\eta^2 = .16$ , the effect size was large and indicated that the self-regulatory pattern accounted for about 16% of the variance in emotional exhaustion. As post hoc analyses (Student–Newman–Keuls) showed, teachers belonging to the H type reported statistically significantly less emotional exhaustion than teachers of the other types. On average, teachers of the U type reported more emotional exhaustion than the H type but less than the two at-risk types. As expected, the A type and the R type scored highest on emotional exhaustion, but they did not differ from each other. In terms of effect sizes, the difference between the average emotional exhaustion of the H type and both the R type (Cohen's  $d = 1.1$ ) and the A type (Cohen's  $d = 0.9$ ) can be classified as large. A large standardized difference was also observable between the U type and the R type (Cohen's  $d = 0.8$ ), and there was a medium-sized difference between the U type and the A type (Cohen's  $d = 0.6$ ). The average mean difference between the H and the U types can be considered small (Cohen's  $d = 0.3$ ; Henson, 2006).

Similar results were found for the job satisfaction of the four teacher types. Overall, we found substantial differences across the four types,  $F(3, 1785) = 59.19, p < .001$ . The moderate effect size ( $\eta^2 = .09$ ) indicated that about 9% of the variance in job satisfaction could be explained by the self-regulatory pattern. Post hoc analyses (Student–Newman–Keuls) showed that the H type scored highest on job satisfaction, followed by the U type. Teachers of the A and R types scored lowest and did not differ from each other. In terms of effect sizes, the difference between the average job satisfaction of the H type and both the R type (Cohen's  $d = 0.8$ ) and the A type (Cohen's  $d = 0.8$ ) was again considerable. A small mean difference was observed between the U type and the R type (Cohen's  $d = 0.4$ ), as well as between the U type and the A

type (Cohen's  $d = 0.4$ ), and between the H type and the U type (Cohen's  $d = 0.3$ ).

Because gender and school track (and age) were not equally distributed across the four types, we performed a further ANCOVA to control for any effects of these variables. Consistent with the homogeneity assumption in ANCOVA, the regression coefficients of the covariates were not found to be statistically significantly different across the different types. As shown in Table 4, the results of the ANCOVA confirmed the findings reported above, and the results remained stable.

### Summary Part 1

In Part 1 of our study, we used latent profile analysis to identify four different self-regulatory types. These empirically derived types were in line with our theoretical expectations of patterns of engagement and resilience. A combination of engagement and resilience was the most adaptive self-regulatory pattern in the school setting, as reflected in the levels of emotional exhaustion and job satisfaction reported. The H type scored more favorably on both well-being indicators, followed by the U type, which in turn differed statistically significantly from the A and R types. The latter two types did not differ statistically significantly from each other. We next investigated whether teachers' self-regulatory styles are also systematically linked to their instructional performance and to students' mathematics achievement and motivational experience in the mathematics classroom.

### Part 2: Resilience, Engagement, and Instructional Performance

Part 2 of our study extended previous research on teachers' instructional behavior in three respects. First, instead of relying on teacher self-reports, we drew on students' ratings of their teachers' instructional quality. Teachers' self-regulatory styles, emotional exhaustion, and job satisfaction are likely to color their perceptions of the quality of their teaching. In fact, a feeling of reduced personal accomplishment is a typical symptom of burnout; this relationship might lead to inflated associations between ratings of teaching quality and stress or emotional exhaustion when teacher

Table 4

*Mean Differences Between the Four Self-Regulatory Types on Emotional Exhaustion and Job Satisfaction: Results of Analysis of Variance (ANOVA) and Analysis of Covariance (ANCOVA)*

Scale	Self-regulatory types				ANOVA		ANCOVA	
	H N = 523	U N = 454	A N = 293	R N = 519	F (3, 1785)	$\eta^2$	F (3, 1785)	$\eta^2$
Emotional exhaustion					116.51*	.16	117.99*	.18
M	1.80 <sub>a</sub>	1.97 <sub>b</sub>	2.36 <sub>c</sub>	2.41 <sub>c</sub>				
SD	0.49	0.58	0.68	0.58				
Job satisfaction					59.19*	.09	60.25*	.10
M	3.24 <sub>a</sub>	3.01 <sub>b</sub>	2.76 <sub>c</sub>	2.72 <sub>c</sub>				
SD	0.59	0.69	0.78	0.70				

*Note.* Age, gender (dummy coded), and school track (dummy coded: 0 = non-academic tracks, 1 = academic track) were included as covariates in the ANCOVA. Means with different subscripts differ statistically significantly in the Student–Newman–Keuls post hoc tests. H = healthy-ambitious type; U = unambitious type; A = excessively ambitious type; R = resigned type.

\*  $p < .0125$ .



self-reports are the only data source. Second, in line with most recent educational research, we took a multidimensional perspective on teachers' classroom behavior. We considered several aspects of instructional quality, including classroom management, tempo, cognitive activation, and perceived social support. Third, we examined whether differences in the instructional performance of teachers are also reflected in students' motivational experience and achievement.

Our main hypothesis concerned differences in the instructional performance of the four teacher types. We expected the different self-regulatory patterns to become manifest in teachers' classroom behavior and therefore observable for students. Specifically, we expected teachers of the H type to show the highest quality of instruction. In contrast, teachers of the R type were not expected to have the necessary resources (in terms of motivational and self-regulatory abilities) to meet the challenges of the classroom and were expected to show particularly poor instructional performance. The highly engaged teachers of the A type were expected to do a much better job in the classroom. The instructional performance of teachers of the U type, who score high on resilience and low on engagement, was a matter of particular interest. Theoretically, they might be more able than their nonresilient counterparts to meet the challenges of the classroom, but it is questionable whether teachers who are not engaged in their profession provide high-quality instruction.

Are self-regulatory patterns also associated with student outcome variables, such as motivational experience and achievement? Provided that teachers of different self-regulatory styles showed the expected differences in instructional performance, we expected to find an association between teachers' self-regulatory types and students' motivational and emotional experience. Based on research on teacher motivation (Skinner & Belmont, 1993), we assumed that an optimal balance of teacher engagement and resilience (H type) would be positively related to students' motivational experience, mediated through teachers' instructional behavior. We further assumed that students taught by teachers of the A type, who were expected to show better performance ratings than teachers low on work engagement (U type and R type), would show higher motivation.

Although it has been established that students' achievement, particularly at the class level, is not easily influenced by short-term interventions and situational features (e.g., Prenzel, 2007), high levels of instructional quality might be expected to enhance student outcomes in the long run. In a more exploratory analysis, we therefore tested whether the four self-regulatory patterns were differentially linked to students' mathematics achievement.

## Method

### Sample

The database for Part 2 of the present study is a subsample of the 1,789 teachers surveyed in Part 1 and comprises those teachers who teach mathematics in each of the 387 ninth-grade classes sampled for PISA 2003 in Germany. The sample size was reduced to 318 from the theoretically possible 387 teachers/classes for several reasons (teachers not consenting to participate, the same teacher teaching more than one PISA class, teachers not completing all instruments, and exclusion of classes for which fewer than

10 student reports were available). Preliminary analyses showed that classes that provided a complete set of teacher and student reports did not differ from classes that did not in terms of school track, mean mathematics achievement, basic cognitive abilities, or mathematics motivation. The teachers' mean age (56.6% male) was 47.9 years ( $SD = 9.0$ ); their teaching experience ranged from 1 to 41 years ( $M = 22.56$ ,  $SD = 9.8$ ). Teachers and students in each class were administered questionnaires assessing teachers' personal characteristics and instructional performance and students' motivation, thus making it possible to link up teacher data with students' reports. Scores on the PISA student achievement tests could also be linked to the teacher data. On average, 12 students completed the questionnaires per teacher (range: 10–18).

To ensure comparability between the teacher sample used in Part 1 ( $N = 1,789$ ) and the subsample used in Part 2 ( $N = 318$ ), we tested whether the two samples differed in terms of age, emotional exhaustion, job satisfaction, the resilience and engagement scales, or the frequency of assignment to the four self-regulatory types. Except for a marginally significant difference in gender, with fewer female teachers in the Part 2 subsample than in the Part 1 sample, no significant differences were found.

### Classification of the Four Types

Our analysis of the subsample of 318 teachers considered in Part 2 was based on the self-regulatory styles to which they had been assigned in Part 1. The frequencies of the four types in the subsample were as follows: 28.6% H, 25.8% U, 15.4% A, and 30.1% R.

### Measures of Instructional Performance

Student ratings of their mathematics teacher's instructional behavior were assessed using a comprehensive instrument developed specifically for the COACTIV study (Kunter et al., 2007). For the purposes of the present study, we considered the teachers' behavior on the four key aspects outlined earlier: classroom management, tempo, cognitive activation, and perceived social support. Previous work has established the factorial validity of the subscales used to assess these four dimensions. We used eight subscales (with three to eight items each) and a 4-point response scale (1 = *disagree*, 4 = *agree*) to assess the four instructional aspects; each dimension was measured by two subscales that were averaged. These four dimensions can be reliably distinguished on the basis of students' ratings (Kunter & Baumert, 2006); moreover, they are associated with several external validation criteria.

As a means of obtaining a score for each teacher on each measure of instructional performance, individual student ratings were aggregated to produce a class mean. Because the following analyses focus on the class level, we first needed to assess the reliability of these aggregated student ratings. The intraclass correlations  $ICC_1$  and  $ICC_2$  are frequently used to determine whether aggregated individual-level ratings are reliable indicators of group-level constructs (Bliese, 2000; Lüdtke, Trautwein, Kunter, & Baumert, 2006). The  $ICC_1$  indicates the proportion of total variance in student ratings that is located between classes, reflecting systematic between-class differences in instructional ratings. The  $ICC_1$  can also be interpreted as an estimate of the reliability of a single student's rating: With growing differences between classes,



individual students' ratings provide increasingly accurate estimates of the class mean. In contrast, the  $ICC_2$  provides an estimate of the reliability of the class mean. It is calculated by applying a Spearman-Brown correction formula—based on the average number of students per class—to the  $ICC_1$ , taking into account the assumption that measurement error decreases with increasing numbers of raters. Class mean ratings can be considered reliable indicators at the class level if student ratings differ systematically between classes (as measured by the  $ICC_1$ ) and there are many raters per class (a fact that informs the  $ICC_2$ ).  $ICC_2$  values above .70 are considered to indicate that the assessment of the class-level construct is sufficiently reliable (Bliese, 2000).

**Classroom management.** This scale taps classroom *disturbance* (sample item: "Students mess around the whole time"; reverse coded; 3 items) and *inefficiency of time use* ("A lot of lesson time is wasted"; reverse coded; 3 items), as rated by students. Higher ratings reflect more efficient classroom management, with low levels of disturbance and efficient time use. The overall scale score can be seen as a reliable indicator of the class level ( $ICC_1 = .30$ ,  $ICC_2 = .84$ ;  $M = 2.46$ ,  $SD = 0.53$ ).

**Tempo.** Two subscales assessed whether the students perceived the pace of classroom interaction to be appropriate. The first subscale measured the perceived tempo of direct *interaction* with the teacher ("Our teacher does not leave us much time to think when asking questions"; 3 items); the second subscale assessed the general tempo of *instruction* ("We move on so quickly in lessons that a lot of students have difficulty keeping up"; 4 items). High scores on these subscales indicate that students perceive the instructional tempo to be too fast ( $ICC_1 = .22$ ,  $ICC_2 = .77$ ;  $M = 2.38$ ,  $SD = 0.38$ ).

**Cognitive activation.** This dimension describes the level of perceived challenge and cognitive autonomy provided by tasks and instruction (Stefanou, Perencevich, DiCintio, & Turner, 2004). Students rated the teacher's use of *cognitively activating tasks* (sample item: "Our teacher changes the setting of problems to find out whether we have really understood the mathematical idea"; 8 items) and the degree of perceived *cognitive autonomy* ("Our teacher lets us use our own strategies to solve difficult problems"; 8 items). Again, we used the mean score on both aspects in the further analyses ( $ICC_1 = .15$ ,  $ICC_2 = .68$ ;  $M = 2.78$ ,  $SD = 0.24$ ).

**Social support.** This social dimension of the teacher's instructional behavior was operationalized by two subscales: the students' perceptions of the teachers' *patience with students' mistakes* ("Our teacher is patient if someone makes a mistake in the lesson"; 3 items) and the teacher's perceived willingness to provide students with *personal support* ("Our teacher soon notices if a student has problems or worries"; 3 items). The mean score reliably assessed students' perceptions of teacher social support at the class level ( $ICC_1 = .30$ ,  $ICC_2 = .84$ ;  $M = 2.76$ ,  $SD = 0.44$ ).

### Assessment of Student Variables

**Students' motivational experience.** On the basis of the idea that positive motivational development depends on the degree to which the social context facilitates the satisfaction of intrinsic needs (Miserandino, 1996; Ryan & Powelson, 1991), we assessed students' perceived competence and autonomy during mathematics lessons as a measure of motivational experience. To this end, students rated perceived teacher support for their needs for autonomy and competence

on seven items from the Intrinsic Need Satisfaction in Class Scale (Kunter, Baumert, & Köller, 2007). All items (e.g., "In mathematics lessons I am considered capable of difficult tasks") were rated on a 4-point Likert scale (1 = *strongly disagree*, 4 = *strongly agree*). The overall scale score can be seen as a reliable indicator on the class level ( $ICC_1 = .13$ ,  $ICC_2 = .70$ ;  $M = 2.53$ ,  $SD = 0.59$ ).

**Students' mathematical achievement.** The international PISA 2003 mathematics assessment provided a measure of students' mathematical achievement. Parameters were estimated on the basis of item response theory scaling, resulting in weighted likelihood estimates ( $M = 0.0$ ,  $SD = 1.0$ ) of each student's mathematical achievement.

## Results

### Self-Regulatory Patterns and Instructional Performance

We expected to find that teachers' self-regulatory patterns were differentially related to their instructional performance, as rated by their students. Four regression analyses were conducted, with classroom management, interaction tempo, cognitive activation, and personal support, respectively, being regressed on teachers' self-regulatory patterns (see Table 5). The categorical variable self-regulatory type was dummy coded, and assignment to the R type was taken as the reference category. Hence, the regression coefficients of the self-regulatory types must be interpreted relative to teachers of the R type.

Results showed that none of the self-regulatory patterns statistically significantly predicted teachers' classroom management as reported by students. However, teachers of the H type scored statistically significantly lower on interaction tempo than teachers of the R type, indicating that their instructional pace was evaluated as being more appropriate to student needs. Neither the U type nor the A type differed statistically significantly from the R type in this respect. About 5% of the between-class variance in ratings of interaction tempo was accounted for by the teacher's self-regulatory pattern ( $R^2 = .05$ ); in terms of effect sizes, this association can be considered to be of small size.

Next, we examined students' reports on the level of cognitive activation provided in mathematics lessons and found statistically significant associations for the H type and the A type. Both the H type and the A type scored higher than the R type on this dimension, indicating that students in classes taught by teachers of the H or A type described their instruction and the tasks set as more cognitively demanding than students in classes taught by teachers of the R type. However, the amount of between-class variance in students' ratings of cognitive activation explained by teachers' self-regulatory patterns was rather small ( $R^2 = .02$ ).

In terms of the personal support the teachers provided, those of the H type again scored higher than teachers of the R type, indicating that engaged and resilient teachers were perceived to be more attentive to their students' individual needs. Teachers of the other two patterns did not differ statistically significantly from the R type. About 8% of the between-class variance in students' ratings of individual teacher support was explained by teacher type ( $R^2 = .08$ ).

Because gender and schools tracks were not equally distributed across the four types, we re-ran the regression analyses, controlling for these variables. As shown in Table 5, the pattern of results remained stable.



Table 5  
Self-Regulatory Types and Instructional Performance: Results of Multiple Regression Analysis

Type/variable	Instructional performance							
	Classroom management		Tempo		Cognitive activation		Personal support	
	M <sub>1</sub> β	M <sub>2</sub> β	M <sub>1</sub> β	M <sub>2</sub> β	M <sub>1</sub> β	M <sub>2</sub> β	M <sub>1</sub> β	M <sub>2</sub> β
Self-regulatory type								
H	.07	−.06	−.20**	−.20**	.17**	.16*	.24**	.23**
U	.04	−.04	.08	.07	.03	.02	−.08	−.08
A	−.03	.04	−.01	−.00	.13*	.11	.07	.08
Control variable								
Teacher gender		−.01		.03		.08		.01
School track		.13*		−.15**		.02		−.14*
R <sup>2</sup>	.01	.02	.05	.08	.02	.04	.08	.09

Note. School track is dummy coded: 1 = academic track, 0 = non-academic tracks. Tempo: low scores indicate an adequate tempo. R (resigned type) is the reference category. M = model; H = healthy-ambitious type; U = unambitious type; A = excessively ambitious type.  
\*  $p < .05$ . \*\*  $p < .01$ .

Self-Regulatory Patterns and Student Outcomes

We further hypothesized that teachers' self-regulatory behaviors were associated with students' motivational experience and achievement. Students' motivational experience scores were regressed on teachers' self-regulatory types. The results (see Table 6) revealed a statistically significant association between the H type and students' motivational experience. Students taught by teachers of the H type reported more pronounced experiences of autonomy and competence than did students taught by teachers of the R type. However, the amount of variance explained in student motivation was small ( $R^2 = .02$ ). Again, the link between teachers' self-regulatory types and student motivation remained stable when we controlled for school track and teacher gender.

Table 6  
Self-Regulatory Types, Instructional Performance, and Students' Motivational Experience: Results of Regression Analyses Testing Mediation Effects

Type/variable	Autonomy/competence			
	M <sub>1</sub> β	M <sub>2</sub> β	M <sub>3</sub> β	M <sub>4</sub> β
Self-regulatory type				
H	.15*	.14*	−.02	−.01
U	.02	.03	.05	.04
A	.02	.03	−.05	−.04
Instructional performance				
Tempo			.07	−.05
Cognitive activation			.26**	.33**
Personal support			.56**	.39**
Control variable				
Teacher gender		.05		.01
School track		−.25**		−.21**
R <sup>2</sup>	.02	.08	.51	.54

Note. School track is dummy coded: 1 = academic track, 0 = non-academic tracks. R (resigned type) is the reference category. M = model; H = healthy-ambitious type; U = unambitious type; A = excessively ambitious type.  
\*  $p < .05$ . \*\*  $p < .01$ .

In an additional step, we tested whether the association found between teachers' self-regulatory type and student motivation was mediated by instructional performance in the classroom, as suggested by our theoretical model. To this end, we used the methodological approach suggested by Baron and Kenny (1986). Following the baseline model (M), in which student motivation was predicted by teachers' self-regulation (see M<sub>1</sub>/M<sub>2</sub> in Table 6), we conducted a multiple regression analysis in which student motivation was regressed on both teachers' self-regulatory type and instructional performance. Because classroom management was not found to be associated with the self-regulatory type (see Table 5), it was not included in this analysis (M<sub>3</sub>). Results revealed that two of the three remaining instructional performance measures were indeed related to the outcome variable: Degree of cognitive activation and personal support provided by the teacher were both statistically significant predictors of students' motivational experience. The tempo of classroom interaction was not related to students' motivational experience. Most important, the association between the H type and students' motivational experience decreased from a significant standardized regression coefficient of .15 when entered by itself to a nonsignificant .02 when instructional performance was included in the regression, suggesting a complete mediation.

In order to investigate which of the two dimensions of instructional performance (cognitive activation vs. personal support) was responsible for the assumed mediation effect, we formally tested the significance of both indirect effects (MacKinnon, Fairchild, & Fritz, 2007), using the product of coefficients strategy, also known as the Sobel test (Baron & Kenny, 1986). Results revealed a significant indirect effect of the H teacher type on students' motivation via perceived personal support ( $z = 4.68, p < .05$ ), as well as an indirect effect via cognitive activation ( $z = 2.87, p > .05$ ), thus suggesting that the association between teachers' self-regulatory type and student motivation is mediated by students' perceptions of both personal support and cognitive activation.

Finally, mathematics achievement scores were regressed on teachers' self-regulatory type. Results showed that students' mathematics achievement on the class level was not related to the

teachers' self-regulatory patterns ( $R^2 = .002$ ), indicating that there was no direct link between teachers' self-regulation and their students' achievement.

### *Summary: Part 2*

In line with our hypotheses, the results revealed statistically significant differences between teachers of different types in student ratings of instructional tempo, cognitive activation, and personal support, with teachers of the H type receiving the most favorable student ratings relative to teachers of the R type. No clear differences were found between teachers of the other self-regulatory types. Teachers' self-regulatory type was also systematically linked to students' motivational experience in lessons, with classes taught by teachers of the H type feeling more competent and autonomous. Congruent with our theoretical framework, this association was mediated by personal support and by the cognitive activation provided by teachers of the H type. No systematic associations were found for students' mathematics achievement. In sum, teachers' self-regulatory styles seem to be an important factor in successful teaching.

### General Discussion

The present research was conducted to investigate how teachers' self-regulatory patterns in response to work-related demands relate to two indicators of their professional success, namely occupational well-being and instructional performance. Our results show that the self-regulatory dimensions of work engagement and resilience are highly relevant in explaining both outcomes. In this section, we discuss the implications of the four self-regulatory types for teachers' occupational well-being and its possible developmental trajectories, as well as for the educational process in terms of instructional performance and educational outcomes.

#### *Patterns of Teacher Self-Regulation: Implications for Occupational Well-Being*

The main objective of our research was to examine individual configurations of two self-regulatory dimensions. Based on conservation of resources theory (Hobfoll, 1989) and the concept of balanced commitment (Hallsten, 1993), we expected the most adaptive self-regulatory style to be characterized by the concurrence of work engagement and resilience. Indeed, teachers of the H type, who exhibited this kind of self-regulatory style, showed the most favorable results on both measures of occupational well-being: They had the lowest ratings on emotional exhaustion and the highest ratings on job satisfaction.

Besides these H-type teachers, we identified three other self-regulatory types, all of which reported lower occupational well-being than the H type. Teachers of the A type and the R type, in particular, can be considered "at risk" in terms of their occupational well-being and, in the long run, probably in terms of their health (Melamed et al., 2006). The outcomes of the R type, who scored lowest on both indicators of occupational well-being, were the least favorable. From the perspective of the conservation of resources theory, it seems reasonable to speculate that, even though teachers of the R type do not invest much energy in their work, their inability to distance themselves from their work prob-

ably poses a further threat to their other resources (e.g., personal resources such as self-esteem, locus of control), which might, in turn, lead to a further decrease in their engagement (Hobfoll & Freedy, 1993). Without intervention measures, it is likely that their well-being will decrease further over time.

Interestingly, although differing fundamentally from the R type in their self-regulatory style, teachers of the A type, who are characterized by high engagement and low resilience, showed similar levels of exhaustion and dissatisfaction as were shown by teachers of the R type. Conversation of resources theory suggests that the resources of teachers of this group are depleted by their excessive exertion (Hobfoll & Freedy, 1993). From an organizational perspective, this result can also be interpreted in the light of the effort-reward imbalance model (Siegrist, 1996), which explains low occupational well-being in terms of an imbalance of the effort exerted and the rewards experienced. Teachers of the A type reported high engagement but received relatively low student ratings (see Part 2 of our study), which may be interpreted as indicating a lack of reward.

Findings on the teachers assigned to the U type, who scored lowest on all engagement subscales but relatively high on resilience, are especially interesting from a motivational perspective. Teachers of this type scored only slightly lower than teachers of the H type on the well-being measures. They were, however, considerably less engaged in their work, which could put them "at risk" of failing to meet the demands of the teaching profession.

These results confirm our main assumption that the abilities to distance oneself emotionally from one's work and to cope with failure (which are high in both the H type and the U type) seem particularly important for teachers' occupational well-being. Given that teachers of the H type reported higher levels of well-being than teachers of the U type, our findings suggest that both resilience against work-related demands and engagement in the teaching profession are important factors in teachers' occupational well-being.

#### *Patterns of Teacher Self-Regulation: Implications for Teachers' Instructional Performance and Student Outcomes*

As shown in Part 2 of our study, teachers' self-regulatory patterns were important not only for their experience of occupational well-being, but also for their instructional performance. Our main hypothesis—that highly engaged and resilient teachers of the H type would outperform teachers of the other types—was confirmed. Indeed, teachers of the H type received the most favorable student ratings in almost all aspects of instructional performance. However, with one exception, the A type was not significantly more successful than the R type in terms of instructional performance. This finding suggests that there is no clear linear association between the engagement shown by a teacher and students' ratings of instructional performance and provides further support for the advantages of achieving a balance between engagement and resilience.

It is interesting that our findings revealed differential associations between the instructional features considered. The strongest effects were found for teachers' personal support and an adequate interaction tempo. A somewhat smaller effect was found for the amount of cognitive activation experienced by students. Classroom



management was the only dimension for which no differences were apparent across teachers of different self-regulatory types. It seems that teachers' ability to perform basic classroom management tasks, such as establishing order, is almost independent of their motivational orientation. Teachers with low engagement were quite able to establish "normal routine" in the classroom (Doyle, 1986), but only teachers of the H type had the capacity to respond adaptively to their students' needs, as reflected in student reports of personal support, interaction tempo, and cognitive activation.

As expected, teachers' self-regulatory patterns not only affected their instructional performance, but were also linked to students' motivational experience. Students in classes taught by engaged and resilient teachers reported more positive motivational experience in mathematics lessons than students in classes taught by any of the other teacher types. Because this effect is mediated by teachers' personal support and, to a lesser degree, by the perceived level of cognitive activation, this result emphasizes that these aspects of instructional performance are crucial for students' motivation. Teachers' self-regulatory patterns did not affect students' mathematics achievement, however. A possible explanation for this finding is that this association is a long-term one that may be mediated by students' motivation and is therefore unlikely to be found in a cross-sectional design. It is also possible that other teacher qualities, such as professional knowledge, have a stronger direct effect on students' achievement (Shulman, 1986). Moreover, the PISA 2003 assessment administered was not tailored to the German Grade 9 mathematics curriculum but instead to tested material covered in Grades 5 through 9. Although the test was implemented at the end of the school year, the current teacher's scope to influence achievement scores for the better was clearly limited.

### Limitations and Future Research

To our knowledge, this is the first study in the area of teacher research to consider how teachers' occupational well-being, students' ratings of instructional performance, and students' motivation and achievement relate to teachers' personal characteristics. At the same time, our study has at least two notable limitations. First, although we report data from a large sample embedded in a large-scale assessment, it remains unclear whether the results can be generalized beyond secondary-level mathematics teachers in Germany. Second, the cross-sectional design of the study does not allow causal inferences to be made between the variables investigated. As mentioned earlier, it would be interesting to investigate the development of self-regulatory processes, occupational well-being, and instructional performance, as well as students' motivational and achievement development over time. In particular, a closer examination of the interplay between—and causal direction of—student characteristics and behavior and teachers' performance would be of major interest (Maslach & Leiter, 1999). Researchers seeking to investigate how classroom experiences influence teachers' performance and occupational well-being need to bear in mind that one teacher interacts with several different classes per day, all of which would probably have to be included in a longitudinal study.

### Practical Implications

In addition to their theoretical relevance, our findings have practical implications, as reflected by the magnitudes of the effect sizes observed. First, teacher education and in-service training that include aspects of teachers' self-regulatory skills and coping behavior might enhance not only teachers' occupational well-being but also their instructional quality. Second, when problems become evident in teachers' occupational well-being or instructional performance, knowledge of the different self-regulatory types might facilitate individually tailored intervention strategies. Some teachers (e.g., the U type) might need more support with motivational aspects, whereas others (e.g., the A type) might benefit more from developing coping strategies to increase their levels of resilience against the potential stressors of the teaching profession.

### References

- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182.
- Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, 9, 3–29.
- Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349–381). San Francisco, CA: Jossey-Bass.
- Bromme, R. (2001). Teacher expertise. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 15459–15465). Oxford, England: Elsevier.
- Collins, A. M., Greeno, J. G., & Resnick, L. B. (2001). Educational learning theory. In N. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 4276–4279). Oxford, England: Elsevier.
- Doyle, W. (1986). Classroom organization and management. In M. Wittrock (Ed.), *Handbook of research on teaching* (pp. 392–431). New York: Macmillan.
- Emmer, E. T., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist*, 36, 103–112.
- Enzmann, D., & Kleiber, D. (1989). *Helfer-Leiden: Stress und Burnout in psychosozialen Berufen* [Stress and burnout in human service professions]. Heidelberg, Germany: Roland Asanger Verlag.
- Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Erlbaum.
- Hackman, J., & Oldham, G. R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology*, 60, 159–170.
- Hakanen, J. J., Bakker, A. B., & Schaufeli, W. B. (2006). Burnout and work engagement among teachers. *Journal of School Psychology*, 42, 495–513.
- Hallberg, U. E., & Schaufeli, W. B. (2006). "Same same" but different? Can work engagement be discriminated from job involvement and organizational commitment? *European Psychologist*, 11, 119–127.
- Hallsten, L. (1993). Burning out: A framework. In W. Schaufeli, C. Maslach, & T. Marek (Eds.), *Professional burnout: Recent developments in theory and research* (pp. 95–113). Philadelphia: Taylor & Francis.
- Henson, R. K. (2006). Effect-size measures and meta-analytic thinking in counseling psychology. *The Counseling Psychologist*, 34, 601–629.
- Hobfoll, S. E. (1989). Conservation of resources: A new attempt at conceptualizing stress. *American Psychologist*, 44, 513–524.



- Hobfoll, S. E., & Freedy, J. (1993). Conservation of resources: A general stress theory applied to burnout. In W. B. Schaufeli, C. Maslach, & T. Marek (Eds.), *Professional burnout: Recent developments in theory and research* (pp. 115–129). Philadelphia: Taylor & Francis.
- Judge, T. A., Thoresen, C. J., Bono, J. E., & Patton, G. K. (2001). The job satisfaction–job performance relationship: A qualitative and quantitative review. *Psychological Bulletin*, 127, 376–407.
- Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research*, 9, 231–251.
- Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction*, 17, 494–509.
- Kunter, M., Klusmann, U., Dubberke, T., Baumert, J., Blum, W., Brunner, M., et al. (2007). Linking aspects of teacher competence to their instruction: Results from the COACTIV project. In M. Prenzel (Ed.), *Studies on the educational quality of schools. The final report on the DFG Priority Programme* (pp. 32–52). Münster, Germany: Waxmann.
- Kyriacou, C. (2001). Teacher stress: Directions for future research. *Educational Review*, 53, 27–35.
- Lazarus, R. S., & Folkman, S. (1987). Transactional theory and research on emotions and coping. *European Journal of Personality*, 1, 141–170.
- Lo, Y., Mendell, N., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, 88, 767–778.
- Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, 10, 21–39.
- Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research*, 9, 215–230.
- MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, 58, 593–614.
- Maslach, C., Jackson, S. E., & Leiter, M. P. (1996). *Maslach Burnout Inventory manual* (3rd ed.). Mountain View, CA: CPP.
- Maslach, C., & Leiter, M. P. (1999). Teacher burnout: A research agenda. In R. Vandenberghe & M. A. Huberman (Eds.), *Understanding and preventing teacher burnout: A sourcebook of international research and practice* (pp. 295–303). Cambridge, England: Cambridge University Press.
- Maslach, C., Schaufeli, W. B., & Leiter, M. P. (2001). Job burnout. *Annual Review of Psychology*, 52, 397–422.
- Melamed, S., Shirom, A., Toker, S., Berliner, S., & Shapira, I. (2006). Burnout and risk of cardiovascular disease: Evidence, possible causal paths, and promising research directions. *Psychological Bulletin*, 132, 327–353.
- Merz, J. (1979). *Berufszufriedenheit von Lehrern: Eine empirische Untersuchung* [Teachers' job satisfaction: An empirical study]. Weinheim, Germany: Beltz.
- Miserandino, M. (1996). Children who do well in school: Individual differences in perceived competence and autonomy in above-average children. *Journal of Educational Psychology*, 88, 203–214.
- Muthén, L. K., & Muthén, B. O. (1998–2004). *Mplus user's guide* (3rd ed.). Los Angeles: Muthén & Muthén.
- Organisation for Economic Cooperation and Development. (2005). *Teachers matter: Attracting, developing and retaining effective teachers*. Paris, France: OECD Publishing.
- Pastor, D. A., Barron, K. E., Miller, B. J., & Davis, S. L. (2007). A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology*, 32, 8–47.
- Prenzel, M. (Ed.). (2007). *Studies on the educational quality of schools. The final report on the DFG Priority Programme*. Münster, Germany: Waxmann.
- Rudow, B. (1999). Stress and burnout in the teaching profession: European studies, issues, and research perspectives. In R. Vandenberghe & M. A. Huberman (Eds.), *Understanding and preventing teacher burnout: A sourcebook of international research and practice* (pp. 38–58). Cambridge, England: Cambridge University Press.
- Ryan, R. M., & Powelson, C. L. (1991). Autonomy and relatedness as fundamental to motivation and education. *Journal of Experimental Education*, 60, 49–66.
- Schaarschmidt, U., Kieschke, U., & Fischer, A. (1999). Beanspruchungsmuster im Lehrerberuf [Patterns of teachers' occupational stress]. *Psychologie in Erziehung und Unterricht*, 46, 244–268.
- Schaufeli, W., & Enzmann, D. (1998). *The burnout companion to study and practice: A critical analysis*. London: Taylor & Francis.
- Semmer, N. (1996). Individual differences, work stress and health. In M. J. Schabracq, J. A. M. Winnubst, & C. L. Cooper (Eds.), *Handbook of work and health psychology* (pp. 53–86). Chichester, England: Wiley.
- Shuell, T. J. (1996). Teaching and learning in a classroom context. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 726–764). New York: Simon & Schuster Macmillan.
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15, 4–14.
- Siegrist, J. (1996). Adverse health effects of high-effort/low-reward conditions. *Journal of Occupational Health Psychology*, 1, 27–41.
- Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology*, 4, 571–581.
- Staub, F. C., & Stern, E. (2002). The nature of teachers' pedagogical content beliefs matters for students' achievement gains: Quasi-experimental evidence from elementary mathematics. *Journal of Educational Psychology*, 94, 344–355.
- Stefanou, K. C., Perencevich, K. C., DiCintio, M., & Turner, J. C. (2004). Supporting autonomy in the classroom: Ways teachers encourage student decision making and ownership. *Educational Psychologist*, 39, 97–110.
- Sternberg, R. J., & Horvath, J. A. (1995). A prototype view of expert teaching. *Educational Researcher*, 24, 9–17.
- Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, 17, 783–805.
- Turner, J. C., Meyer, D. K., Cox, K. E., Logan, C., DiCintio, M., & Thomas, C. T. (1998). Creating contexts for involvement in mathematics. *Journal of Educational Psychology*, 90, 730–745.
- Van Horn, J. E., Taris, T. W., Schaufeli, W. B., & Schreurs, P. J. G. (2004). The structure of occupational well-being: A study among Dutch teachers. *Journal of Occupational and Organizational Psychology*, 77, 365–375.
- Vandenberghe, R., & Huberman, A. M. (Eds.). (1999). *Understanding and preventing teacher burnout: A sourcebook of international research and practice*. Cambridge, England: Cambridge University Press.
- Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. Hagenars & A. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge, England: Cambridge University Press.
- Weinert, F. E., & Helmke, A. (1995). Interclassroom differences in instructional quality and interindividual differences in cognitive development. *Educational Psychologist*, 30, 15–20.
- Weiss, H. M. (2002). Deconstructing job satisfaction: Separating evaluations, beliefs and affective experiences. *Human Resource Management Review*, 22, 173–194.
- Wolpin, J., Burke, R. J., & Greenglass, E. R. (1991). Is job satisfaction an antecedent or a consequence of psychological burnout? *Human Relations*, 44, 193–209.

Received December 12, 2006

Revision received September 6, 2007

Accepted September 10, 2007 ■



# Pedagogical Content Knowledge and Content Knowledge of Secondary Mathematics Teachers

Stefan Krauss, Martin Brunner, Mareike Kunter, and  
Jürgen Baumert  
Max Planck Institute for Human Development

Werner Blum  
University of Kassel

Michael Neubrand  
University of Oldenburg

Alexander Jordan  
University of Bielefeld

Drawing on the work of L. S. Shulman (1986), the authors present a conceptualization of the pedagogical content knowledge and content knowledge of secondary-level mathematics teachers. They describe the theory-based construction of tests to assess these knowledge categories and the implementation of these tests in a sample of German mathematics teachers ( $N = 198$ ). Analyses investigate whether pedagogical content knowledge and content knowledge can be distinguished empirically, and whether the mean level of knowledge and the degree of connectedness between the two knowledge categories depends on mathematical expertise. Findings show that mathematics teachers with an in-depth mathematical training (i.e., teachers qualified to teach at the academic-track Gymnasium) outscore teachers from other school types on both knowledge categories and exhibit a higher degree of cognitive connectedness between the two knowledge categories.

**Keywords:** professional knowledge of mathematics teachers, pedagogical content knowledge, content knowledge, expert teachers

Imagine you are a mathematics teacher. A student puts his hand up and says: "I don't understand why  $-1$  times  $-1$  equals  $+1$ . I know it's the correct result, but it makes no sense to me. Why does multiplying two negative numbers give a positive number?" How would you explain this result to your student? Scenarios like these are typical for the task of teaching. In order to respond appropriately, teachers not only need to understand the mathematical concepts underlying the question, they also need to know how these concepts can best be explained to students.

The relevance of teachers' domain-specific knowledge to high-quality instruction has been discussed, particularly in the context of mathematics teaching (Ball, Lubienski, & Mewborn, 2001;

Fennema & Franke, 1992). Drawing mostly on qualitative data, it has been shown that a deep understanding of mathematical concepts may enable teachers to access a broad repertoire of strategies for explaining and representing mathematical content to their students (Ball, Hill, & Bass, 2005; Ma, 1999). First quantitative evidence shows that students' learning gains in mathematics may be predicted by their teachers' mathematics-related knowledge (Hill, Rowan, & Ball, 2005). The precise nature of teachers' knowledge (i.e., the content and structure of knowledge relating to specific school subjects) remains empirically uncertain, however. Following Shulman (1986, 1987), a theoretical distinction is often drawn between domain-specific subject-matter knowledge, content knowledge (CK), and the knowledge needed for teaching a specific subject, pedagogical content knowledge (PCK). These two knowledge categories may be hypothesized to represent conceptually distinct forms of knowledge, with the former perhaps being the prerequisite for the development of the latter. Alternatively, it is conceivable that the two are merged to form a single body of domain-specific knowledge for teaching (for first results on the knowledge of elementary teachers, see Hill, Schilling, & Ball, 2004, or Phelps & Schilling, 2004). Despite its great relevance to the development of teachers' knowledge and possible implications for teacher training curricula, this issue remains empirically unresolved, primarily because very few instruments are yet available to tap teachers' knowledge directly.

In this article, we present an empirical approach to assessing the knowledge of secondary-level mathematics teachers and investigate whether the theoretical distinction between pedagogical content knowledge and content knowledge can be verified empirically. We report on the theory-driven construction of a test to assess the

---

Stefan Krauss, Martin Brunner, Mareike Kunter, Jürgen Baumert, Center for Educational Research, Max Planck Institute for Human Development, Berlin, Germany; Werner Blum, Department of Mathematics, University of Kassel, Germany; Michael Neubrand, Department of Mathematics, University of Oldenburg, Germany; Alexander Jordan, Department of Mathematics, University of Bielefeld, Germany.

Martin Brunner is now at the Educational Measurement and Applied Cognitive Science (EMACS) Research Unit, University of Luxembourg.

This research, which was part of the larger COACTIV study (see <http://www.mpibberlin.mpg.de/coactiv/index.htm>), was funded by the German Research Foundation (DFG). We thank our colleagues at the Center for Educational Research for their invaluable support with this research project and their comments on previous versions of this manuscript, and Susannah Goss for English editing. We also thank the participating teachers and raters for their support.

Correspondence concerning this article should be addressed to Stefan Krauss, who is now at FB 17, Department of Mathematics, University of Kassel, 34109, Kassel, Germany. E-mail: [skrauss@mathematik.uni-kassel.de](mailto:skrauss@mathematik.uni-kassel.de)

knowledge categories of PCK and CK directly, and on its application in a German teacher sample. Our investigation of commonalities and differences in CK and PCK was informed by the expertise research literature, which has repeatedly found that the knowledge base of experts is not only more extensive than that of novices, but also more connected and integrated (Chi, Feltovich, & Glaser, 1981; Schmidt & Boshuizen, 1992; Simon & Chase, 1973). We therefore investigated whether the association between PCK and CK varies according to the level of mathematical expertise, hypothesizing that the association between the two categories of knowledge will be substantially closer in teachers high in mathematical expertise, possibly even constituting one inseparable body of knowledge. We addressed this question by capitalizing on a quasi-experimental situation that is specific to teacher education in Germany. The amount (and depth) of content-specific instruction provided to teacher candidates varies substantially depending on whether or not they aspire to teach in the academic-track Gymnasium (GY). One group of teachers can therefore be identified as “mathematical content experts.”

### The Concepts of Pedagogical Content Knowledge and Content Knowledge

The distinction between teachers’ knowledge about teaching in a specific domain (PCK) and their domain-specific subject-matter knowledge (CK) has been widely embraced by the research community. The core meaning of pedagogical content knowledge is best represented by Shulman’s (1986, pp. 9–10) original definition, which states that pedagogical content knowledge includes knowledge on how best to represent and formulate the subject to make it comprehensible to others, as well as knowledge on students’ subject-specific conceptions and misconceptions (see also Grossman, 1990). Content knowledge, on the other hand, describes a teacher’s understanding of the structures of his or her domain. According to Shulman, “The teacher need not only understand *that* something is so, the teacher must further understand *why* it is so” (Shulman, 1986, p. 9), which implies that teachers’ content knowledge should represent a deep understanding of the material to be mastered by the students. Shulman’s definitions describe teachers’ pedagogical content knowledge and content knowledge in generic terms. However, domain-specific approaches have been found to provide more valuable, in-depth insights into instructional processes and their prerequisites (Mayer, 2004b).

### Mathematics Teachers’ Pedagogical Content Knowledge and Content Knowledge

Research has identified several aspects that are specifically important to successful mathematics instruction, and that might therefore be used to conceptualize pedagogical content knowledge in a mathematics-specific approach. Most importantly, tasks play a central role in mathematics instruction (Christiansen & Walther, 1986), accounting for much of the time allocated to mathematics lessons. Appropriately selected and implemented mathematical tasks lay the foundations for students’ construction of knowledge and represent powerful learning opportunities (de Corte, Greer, & Verschaffel, 1996; Williams, 2002). Knowledge about the potential of mathematical tasks for learning is thus a first important aspect of mathematical pedagogical content knowledge. Second,

teachers need to work with students’ existing conceptions and prior knowledge. Because errors and mistakes can provide valuable insights into the implicit knowledge of the problem solver (Matz, 1982), it is important for teachers to be aware of typical student misconceptions and comprehension difficulties. Students’ construction of knowledge often only succeeds with instructional support and guidance (Mayer, 2004a), which may entail various forms of explanations or the explicit use of representations. The knowledge of appropriate mathematics-specific instructional methods is thus a third important component of mathematical pedagogical content knowledge. Whereas the latter two components are based directly on Shulman’s (1986) generic conceptualization, we added the component of knowledge on tasks as a third, mathematics-specific component of mathematical pedagogical content knowledge.

The domain-specific conceptualization of teachers’ content knowledge seems straightforward. Clearly, teacher knowledge should go beyond an awareness of the material to be mastered by students; rather, teachers should possess mathematical background knowledge of the content covered in the school curriculum at a much deeper level of understanding than their students. This background knowledge of school curriculum content forms a knowledge base that is specific to teachers, in that it overlaps only partially with the mathematics typically taught at university.

### The Interplay Between Pedagogical Content Knowledge and Content Knowledge

Although the distinction between pedagogical content knowledge and content knowledge appears highly plausible at a theoretical level, its empirical basis is far from certain. Despite repeated calls for reliable and valid measurement assessments of teachers’ knowledge (e.g., Lanahan, Scotchmer, & McLaughlin, 2004), instruments suitable for assessing the categories of teachers’ knowledge remain scarce. To date, only a few studies have investigated PCK and CK empirically. In some of these studies, mathematics teachers’ content knowledge was not tested explicitly, because researchers did not want to give the impression of “testing” teachers (e.g., Kennedy, Ball, & McDiarmid, 1993). Other studies aiming to tap both knowledge categories found that, although it is possible to construct separate tests in principle, the overlap between the two categories was in fact so high that a global factor of knowledge relating to mathematics instruction seemed just as likely (Hill et al., 2004; Kahan, Cooper, & Betha, 2003). At the same time, a study with elementary teachers in the domain of reading found separate categories only, and no common factor (Phelps & Schilling, 2004). One explanation for these inconclusive results may be that the structure of knowledge differs across teacher populations. Studies comparing the knowledge base of experienced and novice teachers (for an overview see Berliner, 2001) suggest that expert teachers not only know more than novice teachers, but that their knowledge is differently structured and may be more highly integrated. This conclusion is in line with findings from expertise research in other domains, which show that experts’ knowledge bases are usually not only more extensive than those of novices, but also more connected and integrated (Chi et al., 1981; Schmidt & Boshuizen, 1992; Simon & Chase, 1973). Whether or not teachers’ pedagogical content knowledge and content knowledge are separable categories of knowledge may therefore be a



function of different levels of expertise. All previous empirical attempts to test models of teachers' knowledge have looked at elementary teachers, who can be assumed to have lower levels of subject-specific expertise. It thus seems worth examining the distinction between CK and PCK in a sample of secondary-level teachers.

### The Present Investigation

The goal of the present article was to construct and to establish a test of secondary mathematics teachers' PCK and CK and to use this test to examine the level and the connectedness of the two knowledge categories in two groups of teachers with different mathematical expertise. To specify our research hypotheses and make them empirically testable, we first clarified our approach to the concepts of *connectedness* and *mathematical expertise*.

Our methodological approach is based on a structural equation framework, in which PCK and CK are conceptualized as latent variables. The latent correlation between PCK and CK is particularly relevant to the issue of distinguishability (see Figure 1a). A higher latent correlation between PCK and CK indicates higher cognitive connectedness between the two knowledge categories. Moreover, a correlation close to one may indicate that the two knowledge constructs indeed form a single, indistinguishable body of knowledge on the cognitive level. In the PCK-CK model (Figure 1a), the means of the two knowledge constructs are depicted by paths from the triangle representing the mean structure to the two latent variables.

We used a quasi-experimental approach to examine whether higher connectedness is a function of higher expertise, attributing expert status on the basis of the teachers' university training (for an overview of alternative ways to identify expert teachers, see Palmer, Stough, Burdinski, & Gonzales, 2005). In Germany, all candidates entering a teacher training program must have graduated from the highest track in the school system, the GY, and received the Abitur qualification. At university, those aspiring to teach at the secondary level must choose between separate degree programs qualifying them to teach either at GY or in the other secondary tracks (e.g., Realschule or Sekundarschule). GY and non-Gymnasium (NGY) teacher education students are usually strictly separated during their university training. One of the main differences in their degree programs is the subject matter covered: Students training to teach at GY cover an in-depth curriculum comparable to that of a master's degree. Relative to their colleagues, who receive less subject-matter training (and usually spend less time at university), GY teachers may therefore be considered mathematical experts, and the two groups of teachers may be contrasted in a quasi-experimental approach. Given that previous studies (e.g., Hill et al., 2004) have presented evidence for a close connection between the two subject-specific knowledge categories, and because CK is often discussed as a prerequisite for PCK, we expected GY teachers would score higher in the PCK test as well, although the difference would probably not be as pronounced as for the CK test.

On the basis of this conceptualization of connectedness and mathematical expertise, we formulated our research hypotheses as follows:

*Hypothesis 1:* GY teachers significantly outscore their colleagues from NGY school types on CK and PCK.

*Hypothesis 2:* The latent correlation between PCK and CK is substantially higher for GY teachers than for NGY teachers (perhaps even approaching  $r = 1$ ).

### Method

#### Participants

The present analyses are based on data obtained from 198 secondary mathematics teachers in Germany. Participants taught mathematics in 10th-grade classes sampled within the framework of a nationally representative student achievement study (cf. Kunter et al., 2007). Thus, our teacher sample can be considered fairly representative of 10th-grade mathematics teachers in Germany. Of the 198 teachers, 85 (55% male) taught at the academic-track GY, and 113 (43% male) at other secondary school types (NGY; e.g., Realschule, Sekundarschule). The average age of participating teachers was 47.2 years ( $SD = 8.4$ ); 46.4 years ( $SD = 9.1$ ; range 28–65) in the GY group and 47.8 years in the NGY group ( $SD = 7.7$ ; range 28–62). Teachers were paid 60 Euro (approximately US \$60 at the time of the survey) for their participation.

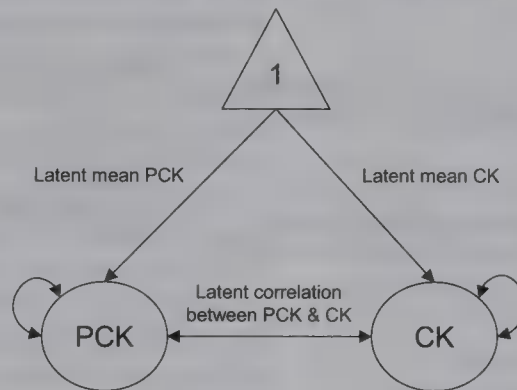
#### Assessment of PCK and CK

*PCK test.* In line with our theoretical framework, the PCK test contained three subscales: knowledge of mathematical tasks (Task), knowledge of student misconceptions and difficulties (Student), and knowledge of mathematics-specific instructional strategies (Instruction). Because the potential of tasks for students' learning can be exploited by considering various solution paths (e.g., Silver, Ghouseini, Gosen, Charalambous, & Strawhun, 2005), we assessed knowledge of tasks by testing teachers' awareness of multiple solution paths: four items required teachers to list as many different ways of solving the task as possible. Knowledge of student misconceptions and difficulties was assessed by presenting teachers with seven scenarios and asking them to detect, analyze (e.g., give cognitive reasons for a comprehension problem), or predict a typical student error or a particular comprehension difficulty. Knowledge of subject-specific instructional strategies was assessed by 10 items requiring teachers to explain mathematical situations (e.g., to provide useful representations, analogies, illustrations, or examples to make mathematical content accessible to students). Sample items for the three PCK subscales (Task, Student, Instruction) are displayed in Figure 2, along with sample responses scoring 1.

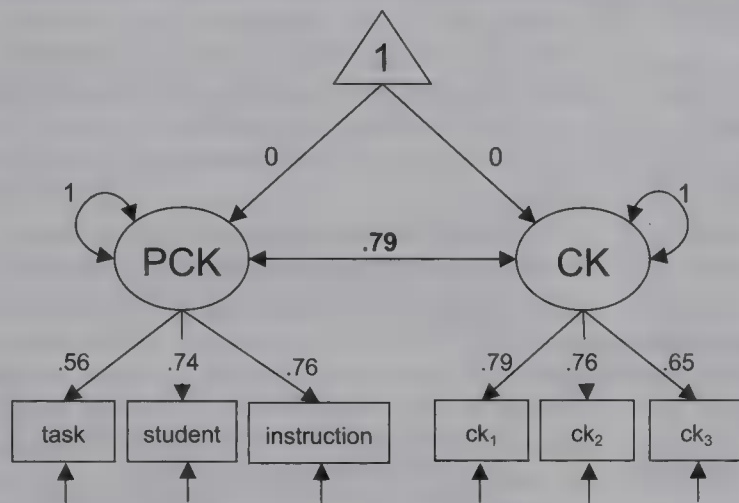
*CK test.* We conceptualized content knowledge as in-depth background knowledge on the contents of the secondary-level mathematics curriculum. Thirteen items were constructed to cover relevant content areas (e.g., arithmetic, algebra, and geometry) and to tap conceptual or procedural skills (Figure 2 presents a sample item).

*Scoring scheme.* All 34 items assessing PCK and CK were open ended. Items with no response or an incorrect response were scored 0; each correct answer was scored 1 (for items requiring several answers, e.g., the multiple solution tasks, the sum of the correct answers was calculated). Each test item was coded by two trained raters independently; in the event of rater disagreement, consensus was reached through discussion. The interrater reliability

## 1a. Structural Conception of PCK and CK



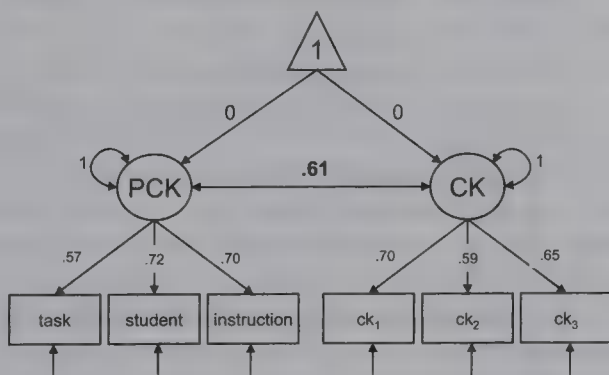
## 1b. Results for the Whole Teacher Sample (N = 198)



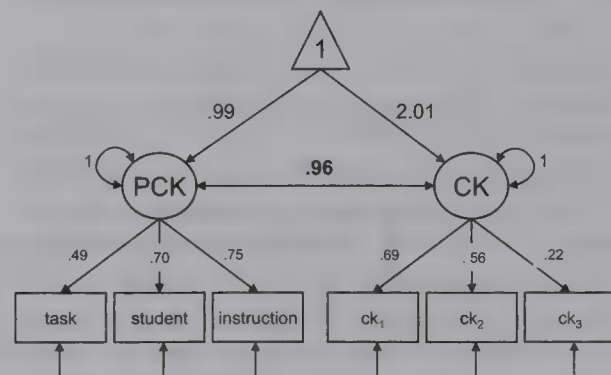
$\chi^2(8, N = 198) = 3.91, p = .87$   
 CFI = 1.00;  
 RMSEA = .00  
 SRMR = .01

## 1c. Results for the Multi-Group Model

Non-Gymnasium teachers  
 (NGY; N = 113)



Gymnasium teachers  
 (GY; N = 85)



$\chi^2(22, N = 198) = 27.13, p = .21$   
 CFI = .98;  
 RMSEA = .05  
 SRMR = .06

*Figure 1.* Model for the latent constructs of PCK and CK: (a) structural conception, (b) results for the whole teacher sample, (c) results for the multigroup model (latent means for the NGY group were set to 0 for purposes of model identification). Model fit indices and standardized model parameters are shown for (b) and (c). SRMR values below .08, RMSEA values below .05, and CFI values above .95 can be considered indicative of a good model fit. PCK = pedagogical content knowledge; CK = content knowledge; GY = Gymnasium; NGY = non-Gymnasium; SRMR = standardized root mean residual; RMSEA = root-mean-square error of approximation; CFI = comparative fit index.



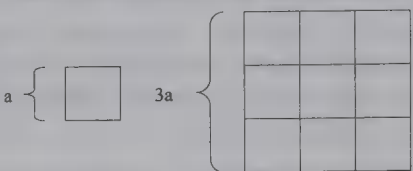
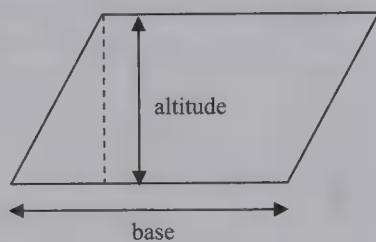


Knowledge Category (Subscale)	Sample Item	Sample response (scoring 1)
PCK Task	<p>How does the surface area of a square change when the side length is tripled? Show your reasoning.</p> <p>Please note down as many different ways of solving this problem (and different reasonings) as possible.</p>	<p><u>Algebraic response</u>            Area of original square: <math>a^2</math>            Area of new square is then <math>(3a)^2 = 9a^2</math>; i.e., 9 times the area of the original square.</p> <p><u>Geometric response</u>            Nine times the area of the original square</p> 
PCK Student	<p>The area of a parallelogram can be calculated by multiplying the length of its base by its altitude.</p>  <p>Please sketch an example of a parallelogram to which students might fail to apply this formula.</p>	 <p>Note: The crucial aspect to be covered in this teacher response is that students might run into problems if the foot of the altitude is outside a given parallelogram.</p>
PCK Instruction	<p>A student says: I don't understand why <math>(-1) \cdot (-1) = 1</math></p> <p>Please outline as many different ways as possible of explaining this mathematical fact to your student.</p>	<p>The "permanence principle," although it does not prove the statement, can be used to illustrate the logic behind the multiplication of two negative numbers and thus foster conceptual understanding:</p> $  \begin{array}{rcl}  3 \cdot (-1) & = & -3 \\  2 \cdot (-1) & = & -2 \\  1 \cdot (-1) & = & -1 \\  0 \cdot (-1) & = & 0 \\  (-1) \cdot (-1) & = & 1 \\  (-2) \cdot (-1) & = & 2  \end{array}  $ <p>Diagram showing a sequence of numbers from 3 down to -2, with arrows indicating the transition from positive to negative and back to positive, illustrating the permanence principle.</p>
CK	<p>Is it true that <math>0.999999... = 1</math> ?</p> <p>Please give detailed reasons for your answer.</p>	<p>Let <math>0.999... = a</math>            Then <math>10a = 9.99...</math>, hence,  <math>10a - a = 9.99... - 0.999...</math></p>  <p>Therefore <math>a = 1</math>; hence, the statement is true</p>

Figure 2. Sample items and corresponding sample responses (scoring 1) from the three subscales of the pedagogical content knowledge (PCK) test (Task, Student, Instruction) and from the content knowledge (CK) test.

ity  $\rho$  (Shavelson & Webb, 1991) was satisfactory (on average,  $\rho$  was .81 with  $SD = .17$ ).

**Procedure.** The assessment of PCK and CK was conducted individually in a separate room at the teacher's school on a workday afternoon. It was administered by a trained test administrator, as a power test with no time constraints. The average time required to complete the 34 items was about 2 hr (approx. 65 min

for the PCK instrument and 55 min for the CK instrument). The teachers were not allowed to use calculators.

#### Statistical Analysis of the PCK-CK Measurement Model

To investigate the structure of knowledge, we employed a confirmatory factor analysis in which the two knowledge catego-

ries were conceptualized as latent constructs based on manifest indicators. The sample size required for structural equation modeling in general and for confirmatory factor analysis in particular has been a matter of some debate. Some scholars recommend a sample size of at least  $N = 200$ ; others argue that no general recommendations can be made because sample size depends strongly on the properties of the model investigated and the data to be analyzed (cf. Jackson, 2003; Marsh, Hau, Balla, & Grayson, 1998). Another way of approaching the issue is to ensure a certain ratio of estimated parameters to participants. Some authors suggest a ratio of at least 1:5 (cf. Marsh et al., 1998). Given the 34 items in our tests (21 measuring PCK and 13 measuring CK), a PCK–CK (item factor) model in which each of the constructs is measured by the respective items would give a ratio of about 1:3, which is unacceptable.

To guarantee that the conclusions derived from our models are valid given our relatively small sample size of  $N = 198$  teachers, we therefore used subscale scores and parcel scores rather than items as manifest indicators (Little, Cunningham, Shahar, & Widaman, 2002), thus reducing the number of parameters to be estimated (e.g., instead of 34, there are just 6 factor loadings; see Figure 1b). This reduction in parameter numbers is particularly relevant for the second part of the analysis, in which the sample is divided into two groups. The latent knowledge construct PCK was measured by the three subscale scores, which were calculated by summing the corresponding item scores. The latent knowledge construct CK was measured by three parcel scores (Little et al., 2002). A preliminary exploratory factor analysis did not identify specific and interpretable subdimensions of CK, but showed that the CK items form one single factor. Therefore, four items with the same stem were assigned to one parcel ( $ck_3$ ) to account for potential task-specific variance that is represented in the corresponding residual term. The remaining nine (unidimensional) items were randomly divided into two further parcels ( $ck_1$  and  $ck_2$ ) following Little et al. (2002).

To tackle the question of empirical distinguishability, we first analyzed the PCK–CK model across all teachers (see Figure 1b), investigating the latent correlation between the two knowledge categories. We then specified a multigroup model (see Figure 1c) to address differences between the two groups of teachers. Although our sample size of  $N = 198$  teachers was sufficient to analyze the PCK–CK model across all teachers (yielding a ratio of about 1:10; see Figure 1b), the multigroup model (Figure 1c) requires closer inspection (here the ratio is about 1:6). Although a ratio of 1:6 seems reasonable, the sample size for the multigroup model is at the lower limit. We therefore investigated the power of the statistical analyses conducted. Table 1 lists the four null hypotheses,  $H_{01}$ – $H_{04}$ , that were central to our statistical analyses.  $H_{01}$  is the null hypothesis stating that PCK and CK do not constitute separable knowledge categories and are not distinguishable in the whole teacher sample.  $H_{02}$  and  $H_{03}$  are the null hypotheses corresponding to our research Hypothesis 1, and  $H_{04}$  is the null hypothesis corresponding to our research Hypothesis 2. Whereas  $H_{01}$  can be seen as a restriction on the model across all teachers (see Figure 1b),  $H_{02}$ – $H_{04}$  can be seen as particular restrictions on the multigroup model (see Figure 1c). Power can be defined as the probability to reject the null hypothesis when it is wrong. Taking this perspective, we followed the methodological recommendations of Satorra and Saris (1985) and calculated within the latent

Table 1  
Power Analysis ( $N = 198$ ;  $\alpha = .05$ )

Null hypotheses to reject	$\Delta\chi^2$	$\Delta df$	Power
$H_{01}$ latent correlation $r_{PCK, CK}$ is 1 in the total teacher sample	20.57	1	0.99
$H_{02}$ latent mean of PCK is identical in both groups of teachers	37.80	1	0.99
$H_{03}$ latent mean of CK is identical in both groups of teachers	106.28	1	1.00
$H_{04}$ latent correlation $r_{PCK, CK}$ is identical in both groups of teachers	6.91	1	0.75

Note. PCK = pedagogical content knowledge; CK = content knowledge.

variable framework the power to reject the null hypotheses  $H_{01}$ – $H_{04}$  given our sample size of  $N = 198$ . To this end, we specified “null models” for  $H_{01}$ – $H_{04}$  (e.g., for  $H_{01}$ , the latent correlation in Figure 1b, was set to 1). The chi-square difference test statistic that results from subtracting the chi-square test statistic of the original model (e.g., the latent correlation in Figure 1b is .79) from the chi-square test statistic of the corresponding “null model” represents the noncentrality parameter that is necessary to calculate statistical power from a noncentral chi-square distribution according to Satorra and Saris (1985). As shown in Table 1, the statistical power to test our hypotheses was acceptable with a sample size of  $N = 198$ .

Before comparing the latent means and correlations of two groups, we had to investigate whether the constructs measured have the same meaning in both groups. In the structural equation literature, it is standard procedure to run a series of invariance tests when doing multigroup comparisons to establish measurement equivalence by showing that (a) a model with factor loadings and intercepts of the manifest variables constrained to be equal across groups fits reasonably well (scalar invariance) and (b) the fit of the scalar invariant model is not much worse than that of a configural invariant or a metric invariant model that imposes fewer equality constraints across groups (Little, 1997; Vandenberg & Lance, 2000). In a configural invariant model, the factorial structure is assumed to be the same across groups; however, factor loadings and intercepts may differ. In the metric invariant model, the factorial structure and factor loadings are assumed to be identical across groups; however, the intercepts may vary. When local misfit is identified in one of these models, the corresponding equality constraint can be relaxed, and a partial invariant model may allow the same conclusions to be drawn as the full invariant model (Byrne, Shavelson, & Muthén, 1989). Our results (see Table 2) show that the fit of the metric invariant model was considerably worse than that of the configural invariant model. However, freeing the factor loading of  $ck_3$  in group GY led to a substantial improvement in model fit (partial metric invariance). The fit of the partial scalar invariant model (without equality constraint on the factor loading of  $ck_3$ ) was worse than that of the partial metric invariant model. The cause of the misfit was the invariance constraint on the intercept of  $ck_3$ . When this intercept was freely estimated in the GY group, the fit of the partial invariant model



Table 2  
*Series of Models Investigating Measurement Equivalence for the Multigroup Model*

Model	$\chi^2$	df	p	CFI	RMSEA	SRMR	$\Delta\chi^2$	$\Delta df$
M1. Configural invariance	15.79	16	.47	1.00	.00	.04		
M2. Metric invariance	36.12	20	.01	0.93	.09	.10		
Difference between M1 and M2							24.14***	4
M3. Partial metric invariance <sup>a</sup>	19.75	19	.41	1.00	.02	.05		
Difference between M2 and M3							57.99***	1
M4. Partial scalar invariance <sup>a</sup>	41.70	23	.01	0.92	.09	.10		
Difference between M4 and M3							25.56***	4
M5. Partial scalar invariance <sup>b</sup>	27.13	22	.21	0.98	.05	.06		
Difference between M5 and M4							33.82***	1

Note. Values for  $\Delta\chi^2$  were calculated according to the formulas provided by B. O. Muthén (1998-2004, p. 22), which generate corrected  $\Delta\chi^2$  statistics when the maximum likelihood estimator MLM is used. CFI = comparative fit index; RMSEA = root-mean-square error of approximation; SRMR = standardized root mean residual.

<sup>a</sup> Factor loading of ck<sub>3</sub> in group Gymnasium freely estimated. <sup>b</sup>Factor loading and intercept of ck<sub>3</sub> in group Gymnasium freely estimated.

\*\*\*  $p < .001$ .

improved and became acceptable (see Table 2). According to Byrne et al. (1989), the partial scalar invariant model, which is presented in Figure 1c, guarantees measurement equivalence; the model can therefore be used to investigate our primary research questions. Furthermore, additional power analyses showed that our sample size yielded a statistical power of at least .99 to distinguish between the models reported in Table 2.

Analyses were run with the Mplus 4.2 software package (L. K. Muthén & Muthén, 1998–2006). Because prior analyses indicated that the distributions of five of the six sum scores (all but the subtest Instruction) deviated from normal distributions (skewness ranged from  $-.65$  to  $.82$ ; kurtosis ranged from  $-1.09$  to  $.07$ ), parameters were computed using the robust maximum likelihood estimator MLM, which is recommended for nonnormally distributed data (L. K. Muthén & Muthén, 1998–2006).

Results

Before we address our core hypotheses on the level and connectedness of PCK and CK in the two groups of teachers, we report the psychometric properties of our test scales and examine the general distinguishability of the two knowledge categories in the whole sample of teachers.

*Psychometric Properties of PCK and CK and Descriptive Information*

To give a first impression of measurement quality, we report the psychometric properties of the overall PCK (21 items) and CK (13 items) scales, which were analyzed by means of parameters derived from classical test theory. Both overall scale scores showed satisfactory reliability, with a Cronbach's alpha of  $\alpha = .77$  for the PCK scale and  $\alpha = .83$  for the CK scale. The items forming each scale discriminated adequately as evident from their (part-whole corrected) item-total correlations (PCK:  $M = .33$ , range:  $.17$  to  $.45$ ; CK:  $M = .48$ ; range:  $.30$  to  $.66$ ). The means and standard deviations for the subscale scores of PCK and the parcel scores of CK are given in Tables 3 and 4, along with the respective intercorrelations. Table 3 shows these descriptives for the whole teacher sample; Table 4, for the GY and NGY groups separately. In sum, these analyses indicate that the newly developed tests succeeded in

reliably assessing the knowledge categories PCK and CK in secondary mathematics teachers.

*Distinguishability of PCK and CK*

Figure 1b presents the results (parameter estimations and fit indices) for the PCK–CK model across all teachers. The fit of the model, as well as the valence and size of the factor loadings, indicates that the structure of teachers' professional knowledge is accurately captured by two latent constructs representing PCK and CK. The latent correlation between PCK and CK was rather high, at  $.79$ , indicating that teachers with higher CK also tend to have higher PCK. To test statistically whether CK and PCK are distinguishable constructs, we estimated a model in which the correlation between CK and PCK was constrained to 1. This led to a statistically significant decline in model fit, as shown by the chi-square difference test,  $\Delta\chi^2(1, N = 198) = 20.57, p < .001$ . Hence, from the perspective of inferential statistics, it can be concluded that, despite their rather high correlational interdependence, CK and PCK represent different constructs, at least when all teachers are considered together.

Table 3  
*Measures of Teachers' Professional Knowledge (Subscales and Parcels): Correlations and Descriptive Statistics for the Total Teacher Sample*

Correlations	Task	Student	Instruction	ck <sub>1</sub>	ck <sub>2</sub>	ck <sub>3</sub>
PCK						
Task (4 items)						
Student (7 items)	.43					
Instruction (10 items)	.43	.56				
CK						
ck <sub>1</sub> (4 items)	.33	.48	.45			
ck <sub>2</sub> (5 items)	.30	.44	.50	.60		
ck <sub>3</sub> (4 items)	.29	.39	.40	.52	.48	
M	6.96	4.93	7.87	1.89	1.47	2.58
SD	2.16	1.98	3.34	1.32	1.27	1.50

Note. Task, Student, and Instruction are the subscales of pedagogical content knowledge (PCK); ck<sub>1</sub>, ck<sub>2</sub>, and ck<sub>3</sub> are the parcels of content knowledge (CK).

Table 4  
*Measures of Teachers' Professional Knowledge (Subscales and Parcels): Correlations and Descriptive Statistics by Teacher Group*

Correlations	Task	Student	Instruction	ck <sub>1</sub>	ck <sub>2</sub>	ck <sub>3</sub>
<b>PCK</b>						
Task		.28	.39	.34	.30	.11
Student	.48		.53	.44	.26	.11
Instruction	.40	.47		.46	.55	.24
<b>CK</b>						
ck <sub>1</sub>	.22	.23	.23		.43	.04
ck <sub>2</sub>	.18	.36	.30	.46		.21
ck <sub>3</sub>	.30	.29	.33	.44	.38	
<b>NGY</b>						
<i>M</i>	6.60	4.27	6.96	1.18	0.96	1.88
<i>SD</i>	2.02	1.85	3.16	0.97	1.02	1.50
<b>GY</b>						
<i>M</i>	7.45	5.81	9.08	2.84	2.16	3.51
<i>SD</i>	2.27	1.80	3.19	1.11	1.23	0.88
<b>Cohen's <i>d</i></b>						
(mean GY vs. NGY)	0.40	0.85	0.67	1.60	1.08	1.28

*Note.* The correlations for Gymnasium (GY) and non-Gymnasium (NGY) teachers are above and below the diagonal, respectively; Task, Student, and Instruction are the subscales of pedagogical content knowledge (PCK); ck<sub>1</sub>, ck<sub>2</sub>, and ck<sub>3</sub> are the parcels of content knowledge (CK); positive *d* values indicate that GY teachers outperformed NGY teachers. According to Cohen (1992), *d* = .3 represents a small effect, *d* = .5 represents a medium effect, and *d* = .8 represents a large effect. All mean differences are significant at *p* < .01.

### *Level of Knowledge and Cognitive Connectedness in the Two Groups of Teachers*

As expected, GY teachers and NGY teachers differed substantially in their mean levels of knowledge (Table 4 gives the means, standard deviations, and intercorrelations of the manifest measures for each teacher group separately). As indicated by Cohen's *d*, the group differences were particularly pronounced in the CK parcels, substantiating our assumption that GY teachers show much deeper mathematical understanding. This finding was further corroborated by the results of the multigroup model, in which PCK and CK were estimated as latent constructs in both groups simultaneously, thus eliminating measurement error (Figure 1c). The results show that GY teachers clearly have a more extensive knowledge base than NGY teachers in both knowledge categories. The latent means of the GY group (PCK: .99, CK: 2.01) were substantially higher than those of the NGY group (fixed to 0 for both constructs; cf. Little, 1997). The corresponding effect sizes were *d* = 1.00 for latent PCK and *d* = 2.15 for latent CK. These differences were statistically significant, as confirmed by two separate chi-square difference tests between the model depicted in Figure 1c and models in which the latent means of CK or PCK, respectively, were constrained to be the same (i.e., 0) across groups (CK:  $\Delta\chi^2[1, N = 198] = 106.28, p < .001$ ; PCK:  $\Delta\chi^2[1, N = 198] = 37.80, p < .001$ ), thus corroborating research Hypothesis 1.

Most important, the multigroup model showed that the two teacher groups differed not only in their knowledge level, but also in the structure of their knowledge base. As Figure 1c shows, the findings also supported research Hypothesis 2, that cognitive connectedness is dependent on the level of mathematical expertise: the latent correla-

tion between CK and PCK was .96 in the GY group and .61 in the NGY group. The statistical significance of the difference between these correlations was again confirmed in a chi-square difference test,  $\Delta\chi^2(1, N = 198) = 6.91, p < .05$ , thus showing a substantially higher degree of cognitive connectedness between the two knowledge categories for teachers in the GY group.

Given the extremely high correlation between PCK and CK in the group of GY teachers, the question arises of whether Shulman's (1986, 1987) two subject-specific knowledge categories are in fact empirically distinguishable in this group of highly knowledgeable teachers. To address this question, we compared the multigroup model (Figure 1c) with a model in which the correlation between the two constructs was fixed to 1 for the GY group. The chi-square difference test between the two nested models was not significant,  $\Delta\chi^2(1, N = 198) = 0.14, p = .72$ , indicating that the two knowledge categories form one body of connected knowledge and that PCK and CK are almost indistinguishable in the group of GY teachers. By contrast, when the correlation between the two constructs was fixed to 1 for the NGY group, the chi-square difference was significant,  $\Delta\chi^2(1, N = 198) = 20.38, p < .001$ , indicating that the two knowledge categories are separate in the group of NGY teachers.

The results thus demonstrate that the two groups of teachers, whose university education differed substantially, differ in both the level and structure of their knowledge. However, we cannot yet rule out the possibility that this finding is simply a manifestation of differences that existed between the groups prior to their teacher training. Although both teacher training tracks have the same formal requirements (Abitur qualification, no entrance exams), it is reasonable to assume that higher achieving students tend to opt to teach at Gymnasium. We investigated this possibility by looking at the teachers' own Abitur grade (which corresponds to the U.S. grade-point average), which was assessed in a biographical questionnaire (the Abitur grade was *z*-standardized for our analyses, with higher values indicating a better grade). The GY teachers indeed had substantially higher Abitur grades than the NGY teachers (NGY: *M* = -.22, *SD* = 1.03; GY: *M* = .26, *SD* = .92; *d* = .49, *p* < .001). To test whether this difference might explain the differences found between the two teacher groups to some degree (Figure 1c), we reran the analyses, with mean differences in the Abitur grade between the two groups of teachers being partialled out of the manifest indicators (subscale scores and parcel scores). The latent means and correlations obtained in the multigroup model were thus adjusted for the mean differences in the Abitur grade. The findings confirmed the pattern of results from our previous analyses: the mean differences found between the two teacher groups were of similar magnitude (standardized latent means of the GY teachers (PCK: .88, *d* = .88; CK: 1.79, *d* = 1.95), as were the latent correlations between PCK and CK (GY: *r* = .94, NGY: *r* = .60). Although the Abitur grade is just one indicator of prior differences, these analyses suggest that, even given the same levels of prior knowledge, it is only in GY teachers that pedagogical content knowledge and content knowledge fuse to form one inseparable category of knowledge.

### *Discussion*

In previous studies, most conclusions about the nature of teachers' knowledge have been drawn using indicators that are rather distal to the concept, such as university grades, number of subject matter courses taken at university (cf. Hill et al., 2005), or questionnaire data on beliefs or subjective theories (cf. Pajares, 1992). Consequently,



numerous calls have been made in the literature for more valid and reliable assessments of teacher knowledge (e.g., Lanahan et al., 2004). In the present study, we constructed and implemented tests to assess the pedagogical content knowledge and content knowledge of secondary mathematics teachers directly. We took a subject-specific approach, thus responding to repeated calls in the literature for general educational psychological theories to be specified for specific school subjects (e.g., Mayer, 2004b); here, for mathematics. Both knowledge categories were measured reliably, the fit for a corresponding structural model was satisfactory, and the mean differences between teachers with different educational backgrounds provided evidence for the empirical validity of the tests.

Our findings provide further evidence for the applicability of Shulman's (1986, 1987) taxonomy of teacher knowledge in empirical settings. More specifically, being informed by literature on expertise research, they offer a possible interpretation for previous inconclusive findings on the distinguishability of CK and PCK (Hill et al., 2004; Kahan et al., 2003; Phelps & Schilling, 2004). Our findings show that the degree of cognitive connectedness between PCK and CK in secondary mathematics teachers is a function of the degree of mathematical expertise. Capitalizing on the quasi-experimental situation of teacher training in Germany, which allowed us to identify teachers with different levels of mathematical expertise, we found that it was not possible to distinguish the two knowledge categories empirically in the high-expertise group of GY teachers, but that this distinction was clearly visible in the group of NGY teachers. Our findings are thus consistent with findings from other domains of expertise research showing that higher expertise often involves stronger integration of different knowledge categories, or "encapsulated knowledge" (Schmidt & Boshuizen, 1992). Consequently, subject-specific knowledge seems to form a common body of expertise in GY teachers, with high levels of CK and PCK alike. Given that GY teacher candidates receive additional in-depth mathematics training at university, but no additional training in teaching mathematics (relative to NGY teacher candidates), their substantially higher PCK scores are remarkable, and may—although very tentatively—be interpreted as a first indication that CK supports the development of PCK.

Practically, our results have at least two implications. First, our instrument might find more widespread application as a psychometric assessment tool that measures teachers' competence directly. In the light of recent developments in the area of teacher education, selection, and accountability—which have raised questions about the competences to be transmitted in teacher education, how schools or districts can evaluate the quality of their teachers, and how to provide teachers with feedback on their strengths and weaknesses—this aspect is of increasing importance. To date, most assessments of teacher quality rely on distal indicators such as university courses, degrees, or grades (Zumwalt & Craig, 2005). Our research identifies another way of gauging teacher qualifications in terms of the assets that seem most important for their primary task of teaching. Due to its pioneer character, our instrument is not yet suitable for use in high-stakes situations that require utmost reliability in identifying different levels of competence. We do not yet know enough about issues such as retest reliability or suitability for other samples, but addressing these questions is an important objective of our ongoing research agenda.

Second, our study provides some valuable insights into the "long arm" of university teacher training. Although our analyses suggest that the two teacher groups probably differed in certain background variables even prior to teacher training, they also

indicate that there must be something specific about either GY teacher training or professional development at GY schools—over and above the different starting levels—that facilitates more extensive knowledge development in this context. Moreover, because no positive correlation was found between years of teaching practice and the two knowledge categories (see Brunner et al., 2006), teacher training can be assumed to be at the core of the development of the two knowledge categories. Future research may be able to provide deeper insights into the acquisition of PCK and CK during teacher training. For instance, longitudinal implementation of our tests at several critical stages in teacher education might provide more accurate information on the timing (e.g., in which phase of teacher education are PCK and CK acquired?) and mechanisms (e.g., which is needed to acquire the other?) of professional expertise development. Such studies may help to create instructional programs (at university and in the classroom) to foster the CK and PCK of student teachers, and to monitor their learning progress with respect to these knowledge categories.

The limitations of our study raise further interesting research questions. First, our study can only provide limited insights into the external validity of our measures. The finding that the teachers with more in-depth training in mathematics scored significantly higher in the content knowledge test may be seen as a first indication of the measure's external validity, but other approaches to the validity issue are also required. For instance, the convergent and discriminant validity of our measures should be investigated by employing our instrument in combination with other direct measures of mathematics teachers' professional knowledge (e.g., the newly developed standardized PRAXIS series; Educational Testing Service, 2006). Even more important, because our knowledge measures were assessed in a standardized testing situation, their implications for authentic learning situations remain to be investigated. Additional research is needed to examine precisely how PCK and CK regulate teaching behavior, and crucially their impact on student learning. Drawing on previous studies (e.g., Fennema & Franke, 1992; Ma, 1999), teachers with higher PCK might be expected to display a broader repertoire of instructional strategies and to be more likely to create cognitively stimulating learning situations. Indeed, first empirical findings indicate that teachers with higher PCK scores on our test tend to set tasks with higher potential for cognitive activation but do not seem to differ from their peers in terms of classroom management (Kunter et al., 2007). These findings provide first evidence for the discriminant validity of our instrument (for the approach of investigating the construct validity of PCK and CK by examining other populations, e.g., subject-matter experts or biology and chemistry teachers, with the tests, see Krauss, Baumert, & Blum, in press). Finally, strictly speaking, the generalizability of our results is limited to secondary mathematics teachers in Germany. Before final conclusions can be drawn about the dimensionality of teacher knowledge, our results need to be replicated in other samples; for instance, with teachers from countries with different educational systems. Last but not least, it is our hope that the present article might not only activate discussion on the professional knowledge of mathematics teachers, but also initiate similar endeavors for other school subjects.

## References

- Ball, D. L., Hill, H. C., & Bass, H. (2005, Fall). Knowing mathematics for teaching. *American Educator*, 14–46.



- Ball, D. L., Lubienski, S. T., & Mewborn, D. S. (2001). Research on teaching mathematics: The unsolved problem of teachers' mathematical knowledge. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 433–456). New York: Macmillan.
- Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research*, 35, 463–482.
- Brunner, M., Kunter, M., Krauss, S., Baumert, J., Blum, W., Dubberke, T., et al. (2006). Welche Zusammenhänge bestehen zwischen dem fachspezifischen Professionswissen von Mathematiklehrkräften und ihrer Ausbildung sowie beruflichen Fortbildung? [How is mathematics teachers' subject-specific professional knowledge related to their initial training and in-service training?] *Zeitschrift für Erziehungswissenschaft*, 4, 521–544.
- Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105, 456–466.
- Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5, 121–152.
- Christiansen, B., & Walther, G. (1986). Task and activity. In B. Christiansen, A. G. Howson, & M. Otte (Eds.), *Perspectives on mathematics education* (pp. 243–307). Dordrecht, the Netherlands: Reidel.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159.
- de Corte, E., Greer, B., & Verschaffel, L. (1996). Mathematics teaching and learning. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 491–549). New York: Macmillan.
- Educational Testing Service. (2006). *Proper use of the Praxis series and related assessments*. Retrieved June 19, 2006, from <http://www.ets.org/Media/Tests/PRAXIS/pdf/guidelines.pdf>
- Fennema, E., & Franke, M. L. (1992). Teachers' knowledge and its impact. In D. A. Grouws (Ed.), *NCTM Handbook of research on mathematics teaching and learning* (pp. 147–164). New York: Macmillan.
- Grossman, P. L. (1990). *The making of a teacher: Teacher knowledge and teacher education*. New York: Teachers College Press.
- Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, 42, 371–406.
- Hill, H. C., Schilling, S. G., & Ball, D. L. (2004). Developing measures of teachers' mathematics knowledge for teaching. *Elementary School Journal*, 105, 11–30.
- Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling*, 10, 128–141.
- Kahan, J. A., Cooper, D. A., & Betha, K. A. (2003). The role of mathematics teachers' content knowledge in their teaching: A framework for research applied to a study of student teachers. *Journal of Mathematics Teacher Education*, 6, 223–252.
- Kennedy, M. M., Ball, D. L., & McDiarmid, G. W. (1993). *A study package for examining and tracking changes in teachers' knowledge*. East Lansing, MI: The National Center for Research on Teacher Education/Michigan State University.
- Krauss, S., Baumert, J., & Blum, W. (in press). Secondary mathematics teachers' pedagogical content knowledge and content knowledge: Validation of the COACTIV constructs. *The International Journal on Mathematics Education*.
- Kunter, M., Klusmann, U., Dubberke, T., Baumert, J., Blum, W., Brunner, M., et al. (2007). Linking aspects of teacher competence to their instruction: Results from the COACTIV project. In M. Prenzel (Ed.), *Studies on the educational quality of schools. The final report on the DFG Priority Programme* (pp. 39–59). Münster, Germany: Waxmann.
- Lanahan, L., Scotchmer, M., & McLaughlin, M. (2004). *Methodological critique of current NCES survey measures of instructional processes*. Retrieved May 14, 2006, from [http://www.air.org/news\\_events/documents/AERA2004NCESMeasures.pdf](http://www.air.org/news_events/documents/AERA2004NCESMeasures.pdf)
- Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, 32, 53–76.
- Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighting the merits. *Structural Equation Modeling*, 9, 151–173.
- Ma, L. (1999). *Knowing and Teaching Elementary Mathematics. Teachers' Understanding of Fundamental Mathematics in China and the United States*. Mahwah, NJ: Erlbaum.
- Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, 33, 181–220.
- Matz, M. (1982). Towards a process model for high school algebra errors. In D. Sleeman & J. S. Brown (Eds.), *Intelligent tutoring systems* (pp. 25–50). London: Academic Press.
- Mayer, R. E. (2004a). Should there be a three-strikes rule against pure discovery learning? *American Psychologist*, 59, 14–19.
- Mayer, R. E. (2004b). Teaching of subject matter. *Annual Review of Psychology*, 55, 715–744.
- Muthén, B. O. (1998–2004). *Mplus technical appendices*. Los Angeles, CA: Muthén & Muthén.
- Muthén, L. K., & Muthén, B. O. (1998–2006). *Mplus user's guide* (4th ed.). Los Angeles, CA: Muthén & Muthén.
- Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research*, 62, 307–332.
- Palmer, D. J., Stough, L. M., Burdinski, T. K., Jr., & Gonzales, M. (2005). Identifying teacher expertise: An examination of researchers' decision making. *Educational Psychologist*, 40, 13–25.
- Phelps, G., & Schilling, S. (2004). Developing measures of content knowledge for teaching reading. *Elementary School Journal*, 105, 31–48.
- Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, 50, 83–90.
- Schmidt, H. G., & Boshuizen, H. P. A. (1992). Encapsulation of biomedical knowledge. In D. A. Evans & V. L. Patel (Eds.), *Advanced models of cognition for medical training and practice* (pp. 265–282). New York: Springer.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15, 4–14.
- Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57, 1–22.
- Silver, E. A., Ghouseini, H., Gosen, D., Charalambous, C., & Strawhun, B. (2005). Moving from rhetoric to praxis: Issues faced by teachers in having students consider multiple solutions for problems in the mathematics classroom. *Journal of Mathematical Behavior*, 24, 287–301.
- Simon, H. A., & Chase, W. G. (1973). Skill in chess. *American Scientist*, 61, 394–403.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3, 4–70.
- Williams, G. (2002). Identifying tasks that promote creative thinking in mathematics: A tool. In B. Barton, K. Irwin, M. Pfannkuch, & M. Thomas (Eds.), *Mathematical education in the South Pacific: Proceedings of the 25th Annual Conference of the Mathematics Education Research Group of Australasia: Vol. II* (pp. 698–705). Auckland, New Zealand: MERGA.
- Zumwalt, K., & Craig, E. (2005). Teachers' Characteristics: Research on the Indicators of Quality. In M. Cochran-Smith & K. M. Zeichner (Eds.), *Studying Teacher Education: The Report of the AERA Panel on Research and Teacher Education* (pp. 157–260). Mahwah, NJ: Erlbaum.

Received August 29, 2006

Revision received October 4, 2007

Accepted October 7, 2007 ■



# Instructions to Authors

## Journal of Educational Psychology

www.apa.org/journals/edu

The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

**Manuscript preparation.** Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (5th ed.). Manuscripts may be copyedited for bias-free language (see chap. 2 of the *Publication Manual*). Formatting instructions (all copy must be double-spaced) and instructions on the preparation of tables, figures, references, metrics, and abstracts appear in the *Manual*. For APA's Checklist for Manuscript Submission, see [www.apa.org/journals](http://www.apa.org/journals). **Abstract and keywords.** All manuscripts must include an abstract containing a maximum of 180 words typed on a separate page. After the abstract, please supply up to five keywords or brief phrases. **References.** References should be listed in alphabetical order. Each listed reference should be cited in text, and each text citation should be listed in the References. Basic formats are as follows:

Haag, L., & Stern, E. (2003). In search of the benefits of learning Latin. *Journal of Educational Psychology*, 95, 174–178.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Johnson, D. W., & Johnson, R. T. (1990). Cooperative learning and achievement. In S. Sharon (Ed.), *Cooperative learning: Theory and research* (pp. 173–202). New York: Praeger.

Adequate description of participants is critical to the science and practice of educational psychology; this allows readers to assess the results, determine generalizability of findings, and make comparisons in replications, extensions, literature reviews, or secondary data analyses. Authors should see guidelines for sample–subject description in the *Manual*. Appropriate indexes of effect size or strength of relationship should be incorporated in the results section of the manuscript (see pp. 5, 25–26 of the *Manual*). Information that allows the reader to assess not only the significance but also the magnitude of the observed effects or relationships clarifies the importance of the findings. **Figures.** Graphics files are welcome if supplied as Tiff, EPS, or PowerPoint. High-quality printouts or glossies are needed for *all* figures. The minimum line weight for line art is 0.5 point for optimal printing. When possible, please place symbol legends below the figure image instead of to the side. Original color figures can be printed in color at the editor's and publisher's discretion provided the author agrees to pay \$255 for one figure, \$425 for two figures, \$575 for three figures, \$675 for four figures, and \$55 for each additional figure.

**Publication policies.** APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. APA policy regarding posting articles on the Internet may be found at [www.apa.org/journals](http://www.apa.org/journals). In addition, it is a violation of APA Ethical Principles to publish "as original data, data that have been previously published" (Standard 8.13). As this is a primary journal that publishes original material only, APA policy prohibits publication of any manuscript or data that have already been published in whole or substantial part elsewhere. Authors have an obligation to consult journal editors concerning prior publication of any data on which their article depends. In addition, APA Ethical Principles specify that "after research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use

such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release" (Standard 8.14). Authors must have available their data throughout the editorial review process and for at least 5 years after the date of publication.

**Masked review policy.** The *Journal* has a masked review policy, which means that the identities of both authors and reviewers are masked. Every effort should be made by the authors to see that the manuscript itself contains no clues to their identities. Authors should never use first person (*I, my, we, our*) when referring to a study conducted by the author(s) or when doing so reveals the authors' identities, e.g., "in our previous work, Johnson et al., 1998 reported that . . ." Instead, references to the authors' work should be in third person, e.g., "Johnson et al. (1998) reported that . . ." The authors' institutional affiliations should also be masked in the manuscript. Authors submitting manuscripts are required to include in the cover letter the title of the manuscript along with all authors' names and institutional affiliations. However, the first page of the manuscript should omit the authors' names and affiliations, but should include the title of the manuscript and the date it is submitted. Responsibility for masking the manuscript rests with the authors; manuscripts will be returned to the author if not appropriately masked. If the manuscript is accepted, authors will be asked to make changes in wording so that the paper is no longer masked. Authors are required to state in writing that they have complied with APA ethical standards in the treatment of their sample, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained at [www.apa.org/ethics/](http://www.apa.org/ethics/) or by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242. APA requires authors to reveal any possible conflict of interest in the conduct and reporting of research (e.g., financial interests in a test procedure, funding by pharmaceutical companies for drug research). Authors of accepted manuscripts will be required to transfer copyright to APA.

**Permissions.** Authors of accepted papers are required to obtain and provide to the editor on final acceptance all necessary permissions to reproduce in print and electronic form any copyrighted work, including, for example, test materials or portions thereof and photographs of people.

**Supplemental materials.** APA can now place supplementary materials online, which will be available via the published article in the PsycARTICLES database. To submit such materials, please see [www.apa.org/journals/authors/suppmaterial.html](http://www.apa.org/journals/authors/suppmaterial.html) for details. Authors of accepted papers will be asked to work with the editor and production staff to provide supplementary materials as appropriate.

**Submission.** Authors should submit their manuscripts electronically via the Manuscript Submission Portal at [www.apa.org/journals/edu](http://www.apa.org/journals/edu) (follow the link "Submit Manuscripts Electronically"). A checklist for manuscript submission, including guidelines for preparing the electronic file, can be found at [www.apa.org/journals/](http://www.apa.org/journals/). Correspondence regarding manuscripts should be sent to the Editor, Art Graessen, University of Memphis, Journal of Educational Psychology, 202 Psychology Building, Memphis, TN 38152-3230. In addition to addresses and phone numbers, authors should supply e-mail addresses, as most communications will be by e-mail. Fax numbers, if available, should also be provided for potential use by the editorial office and later by the production office. Authors should keep a copy of the manuscript to guard against loss. E-mail correspondence may be addressed to [jedgar@memphis.edu](mailto:jedgar@memphis.edu).

**Preparing files for production.** If your manuscript is accepted for publication, please follow the guidelines for file formats and naming provided at [www.apa.org/journals/authors/preparing\\_efiles.html](http://www.apa.org/journals/authors/preparing_efiles.html). If your manuscript was mask reviewed, please ensure that the final version for production includes a byline and full author note for typesetting.



# WORKING WITH PARENTS OF NONCOMPLIANT CHILDREN

## A GUIDE TO EVIDENCE-BASED PARENT TRAINING FOR PRACTITIONERS AND STUDENTS

**Mark D. Shriver and Keith D. Allen**

**T**his book presents an in-depth look at evidence-based programs for training parents of children with behavior problems. Authors Mark D. Shriver and Keith D. Allen review the empirical support for four major programs, as well as some more popular programs that lack strong empirical support. Throughout this review they teach readers how to identify the best research in parent training, how to prepare for parent training sessions, and finally show how to translate this research into everyday practice.

Parents usually are the most influential people in a child's life. So when child problems like noncompliance, hyperactivity, sleep problems, feeding problems, toileting problems, academic problems, and adolescent-parent conflict arise it is important for parents to take on the primary role in their resolution. This book explains *what* parents are taught when collaborating with a psychologist or counselor and *how* they are taught effectively. Practitioners, whether they are experienced, new to parent training, or students of the field, will find this book to be a valuable resource for taking interventions developed under tightly controlled research conditions and adapting them to the conditions of day-to-day practice, where resources are more limited and presenting problems are often more complex. 2008. 280 pages. Hardcover.

**Series: Applying Psychology to the Schools (Division 16)**

**CONTENTS:** Preface ■ Introduction ■ **Part I: Searching for the Best Available Evidence** ■ Chapter 1. Parenting and Parent Training ■ Chapter 2. Empirically Supported Parent Training Programs ■ Chapter 3. Evaluating the Scientific Merit of Parent Training Alternatives ■ **Part II: Developing Clinical Expertise** ■ Chapter 4. Conceptual Foundations of the Empirically Supported Parent Training Programs ■ Chapter 5. How to Teach Parents ■ Chapter 6. Cultural Issues in Parent Training ■ **Part III: Integrating and Translating Research Into Everyday Practice** ■ Chapter 7. Beyond Noncompliance: Developing Evidence-Based Parent Training Interventions ■ Chapter 8. Delivering Evidence-Based Parent Training: From Research to Practice ■ Chapter 9. Parent Training: Prevention and Future Research ■ References

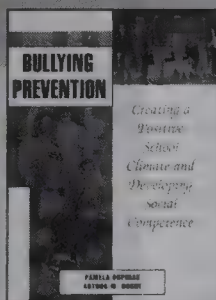
**ISBN 978-1-4338-0344-4 ■ Item # 4317155 ■ List: \$59.95 ■ APA Member/Affiliate: \$49.95**

# Working With Parents of Noncompliant Children

A Guide to Evidence-Based Parent  
Training for Practitioners and Students



Mark D. Shriver and Keith D. Allen



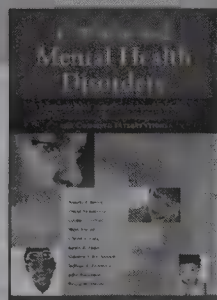
### BULLYING PREVENTION

**Creating ■ Positive School Climate and Developing Social Competence**

Pamela Orpinas and Arthur M. Horne

2006 ■ 293 pages ■ Hardcover

ISBN 978-1-59147-282-7 ■ Item # 4317082 ■ List: \$59.95 ■ APA Member/Affiliate: \$49.95



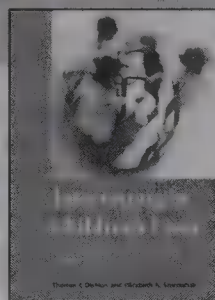
### CHILDHOOD MENTAL HEALTH DISORDERS

**Evidence Base and Contextual Factors for Psychosocial,  
Psychopharmacological, and Combined Interventions**

Ronald T. Brown, et al.

2008 ■ 207 pages ■ Hardcover

ISBN 978-1-4338-0170-9 ■ Item # 4317137 ■ List: \$59.95 ■ APA Member/Affiliate: \$49.95



### INTERVENING IN CHILDREN'S LIVES

**An Ecological, Family-Centered Approach to Mental Health Care**

Thomas J. Dishion and Elizabeth A. Stormshak

2007 ■ 320 pages ■ Hardcover

ISBN 978-1-59147-428-9 ■ Item # 4317115 ■ List: \$69.95 ■ APA Member/Affiliate: \$49.95

**ALSO AVAILABLE**

## APA Books

Ordering Information

**800-374-2721**

[www.apa.org/books](http://www.apa.org/books)

In Washington, DC,

call: 202-336-5510

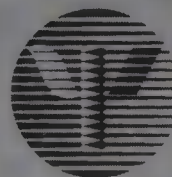
TDD/TTY: 202-336-6123

Fax: 202-336-5502

In Europe, Africa, or the Middle East,

call: 44-207-240-0856

AD0584



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION



# BEST SELLERS

from the American Psychological Association



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

## Handbook of Bereavement Research and Practice

Advances in Theory and Intervention

*Edited by Margaret Stroebe,  
Robert O. Hansson, Henk Schut,  
and Wolfgang Stroebe*

2008. 624 pages. Hardcover.

List: \$69.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-351-2 • Item # 4318045

## Cultural Competence in Trauma Therapy Beyond the Flashback

*Laura S. Brown*

2008. 280 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0337-6 • Item # 4317149

## Working With Parents of Noncompliant Children

A Guide to Evidence-Based  
Parent Training for Practitioners  
and Students

*Mark D. Shriver and Keith D. Allen*

2008. 280 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0344-4 • Item # 4317155

## Transcending Self-Interest Psychological Explorations of the Quiet Ego

*Edited by Heidi A. Wayment  
and Jack J. Bauer*

2008. 272 pages. Hardcover.

List: \$79.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0340-6 • Item # 4317153

## Women Street Hustlers Who They Are and How They Survive

*Barbara A. Rockell*

2008. 232 pages. Hardcover.

List: \$69.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0333-8 • Item # 4316104

## Law and Mental Health Professionals Massachusetts

Third Edition

*Justice Jonathan Brant*

2008. 288 pages. Hardcover.

List: \$99.95

APA Member/Affiliate: \$74.95

ISBN 978-1-4338-0334-5 • Item # 4315011

## Courtroom Modifications for Child Witnesses

Law and Science

*Susan R. Hall and Bruce D. Sales*

2008. 368 pages. Hardcover.

List: \$79.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0354-3 • Item # 4317156

## Undergraduate Writing in Psychology

Learning to Tell the Scientific Story

*R. Eric Landrum*

2008. 208 pages. Paperback.

List: \$29.95

APA Member/Affiliate: \$24.95

ISBN 978-1-4338-0332-1 • Item # 4313015

## Applying to Graduate School in Psychology

Advice From Successful Students  
and Prominent Psychologists

*Edited by Amanda C. Kracen  
and Ian J. Wallace*

2008. 304 pages. Paperback.

List: \$34.95

APA Member/Affiliate: \$29.95

ISBN 978-1-4338-0345-1 • Item # 4313018

## Studying Psychology in the United States

Expert Guidance  
for International Students

*Edited by Nadia T. Hasan, Nadya A. Fouad,  
and Carol Williams-Nickelson*

2008. 208 pages. Paperback.

List: \$29.95

APA Member/Affiliate: \$29.95

ISBN 978-1-4338-0341-3 • Item # 4313017

## Surviving Graduate School in Psychology

A Pocket Mentor

*Tara L. Kuther*

2008. 344 pages. Paperback.

List: \$34.95

APA Member/Affiliate: \$29.95

ISBN 978-1-4338-0346-8 • Item # 4313019

## Psychology ■ ■ Major

Is It Right for Me and What Can I Do  
With My Degree?

*Donna E. Palladino Schultheiss*

2008. 288 pages. Paperback.

List: \$29.95

APA Member/Affiliate: \$24.95

ISBN 978-1-4338-0336-9 • Item # 4313016

## Internships in Psychology

The APAGS Workbook for  
Writing Successful Applications  
and Finding the Right Fit

Second Edition

*Carol Williams-Nickelson,  
Mitchell J. Prinstein, and W. Gregory Keilin*

2008. 144 pages. Paperback.

List: \$24.95

APA Member/Affiliate: \$19.95

ISBN 978-1-4338-0355-0 • Item # 4313021

## Favorite Activities for the Teaching of Psychology

*Edited by Ludy T. Benjamin, Jr.*

2008. 400 pages. Paperback.

List: \$34.95

APA Member/Affiliate: \$29.95

ISBN 978-1-4338-0349-9 • Item # 4316105

## Ethical Conflicts in Psychology

Fourth Edition

*Donald N. Bersoff*

2008. 632 pages.

Hardcover.

List: \$69.95

APA Member/Affiliate: \$54.95

ISBN 978-1-4338-0350-5 • Item # 4312009

Paperback.

List: \$49.95

APA Member/Affiliate: \$39.95

ISBN 978-1-4338-0353-6 • Item # 4312012

## Ethics Desk Reference for Psychologists

*Jeffrey E. Barnett and W. Brad Johnson*

2008. 200 pages. Spiral Bound.

List: \$39.95

APA Member/Affiliate: \$34.95

ISBN 978-1-4338-0352-9 • Item # 4312011

## Casebook for Clinical Supervision

■ Competency-Based Approach

*Edited by Carol A. Falender  
and Edward P. Shafranske*

2008. 272 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0342-0 • Item # 4317154

## The Collaborative Psychotherapist

Creating Reciprocal Relationships  
With Medical Professionals

*Nancy Breen Ruddy, Dorothy A. Borresen,  
and William B. Gunn*

2008. 232 pages. Hardcover.

List: \$49.95

APA Member/Affiliate: \$39.95

ISBN 978-1-4338-0338-3 • Item # 4317152

## Psychotherapy With Cardiac Patients Behavioral Cardiology in Practice

*Ellen A. Dornelas*

2008 pages. 280 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0356-7 • Item # 4317157

## Dieting, Overweight, and Obesity

Self-Regulation in  
■ Food-Rich Environment

*Wolfgang Stroebe*

2008. 256 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0335-2 • Item # 4317148

## The I/O Consultant

Advice and Insights for  
Building ■ Successful Career

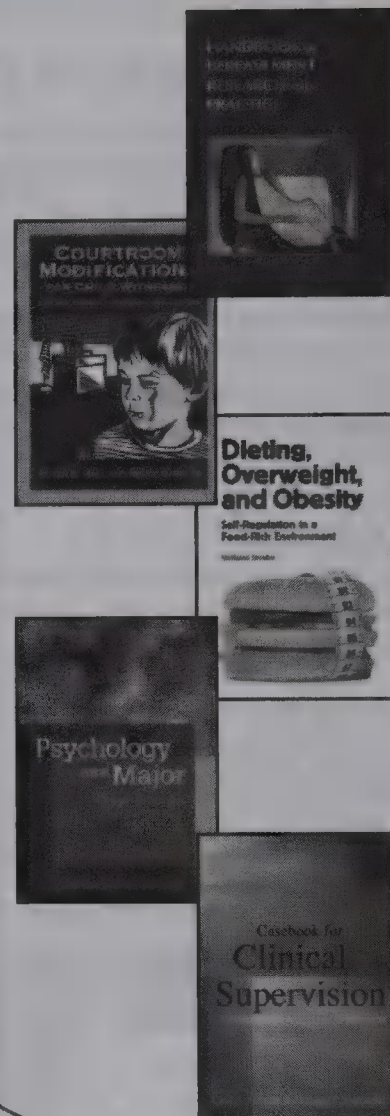
*Edited by Jerry W. Hedge  
and Walter C. Borman*

2008. 328 pages. Hardcover.

List: \$79.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0339-0 • Item # 4316106



To Order: 800-374-2721 • [www.apa.org/books](http://www.apa.org/books)

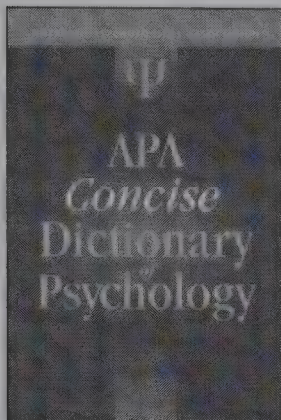


# NEW RELEASES

from the American Psychological Association



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

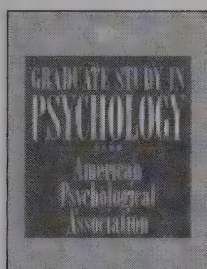


## APA Concise Dictionary of Psychology

This abridged version of the landmark, critically acclaimed *APA Dictionary of Psychology* maintains all the best features of the parent dictionary, including

- 10,000 entries offering clear and authoritative definitions—including many revised and updated from the parent dictionary
- Balanced coverage of over 90 subareas across the field of psychology—including clinical, experimental, neuropsychology, cognitive, personality, social, developmental, health, research methodology, and many others
- Thousands of incisive cross-references that deepen the user's understanding of related topics
- A "Quick Guide to Use" that explains stylistic and formal features at a glance
- Appendixes listing major figures in the history of psychology and psychological therapies and approaches

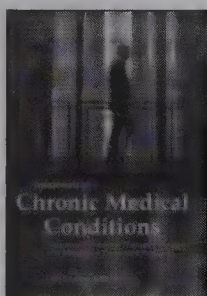
2009. 596 pages. Hardcover.  
ISBN 978-1-4338-0391-8; Item # 4311009  
List: \$39.95; APA Member/Affiliate: \$39.95



## Graduate Study in Psychology

2009 EDITION

2009. 832 pages. Paperback.  
ISBN 978-1-4338-0395-6; Item # 4270092  
List: \$27.95; APA Member/Affiliate: \$22.95

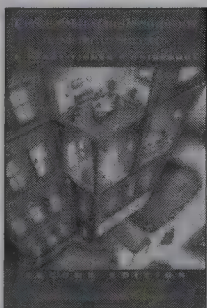


## Treatment of Chronic Medical Conditions

Cognitive-Behavioral Therapy  
Strategies and Integrative  
Treatment Protocols

*Len Sperry*

2009. 330 pages. Hardcover.  
ISBN 978-1-4338-0389-5; Item # 4317164  
List: \$59.95; APA Member/Affiliate: \$49.95

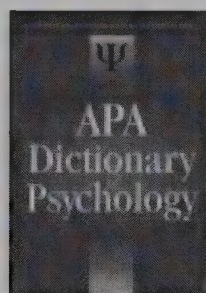


## Clinical Health Psychology in Medical Settings

A Practitioner's Guidebook

SECOND EDITION

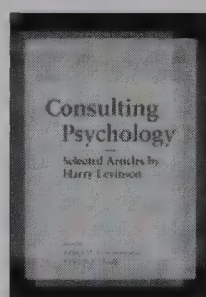
Cynthia D. Belar and William W. Deardorff  
2009. 358 pages. Hardcover.  
ISBN 978-1-4338-0378-9; Item # 4317163  
List: \$59.95; APA Member/Affiliate: \$49.95



Bestseller!

## APA Dictionary of Psychology

2007. 1,024 pages. Hardcover.  
ISBN 978-1-59147-380-0; Item # 4311007  
List: \$59.95; APA member/Affiliate: \$49.95

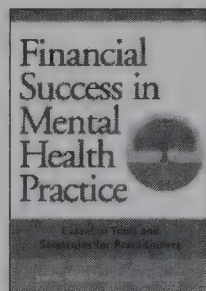


## Consulting Psychology

Selected Articles  
by Harry Levinson

*Edited by Arthur M. Freedman  
and Kenneth H. Bradt*

2009. 453 pages. Hardcover.  
ISBN 978-1-4338-0376-5; Item # 4316107  
List: \$59.95; APA Member/Affiliate: \$49.95



## Financial Success in Mental Health Practice

Essential Tools and Strategies  
for Practitioners

*Steven Walfish and Jeffrey E. Barnett*

2009. 344 pages. Hardcover.  
ISBN 978-1-4338-0374-1; Item # 4317162  
List: \$59.95; APA Member/Affiliate: \$49.95

AD0600

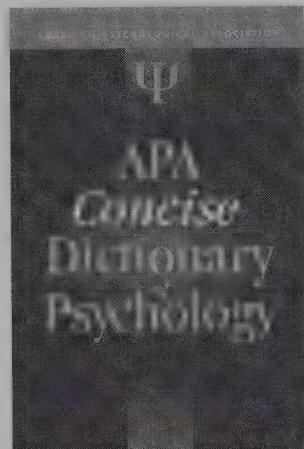
To Order: 800-374-2721 • [www.apa.org/books](http://www.apa.org/books)





## SPECIAL INTRODUCTORY TIME-LIMITED OFFER!

Order online to receive the *APA Concise Dictionary of Psychology* for only **\$29.95!**



### APA Concise Dictionary of Psychology

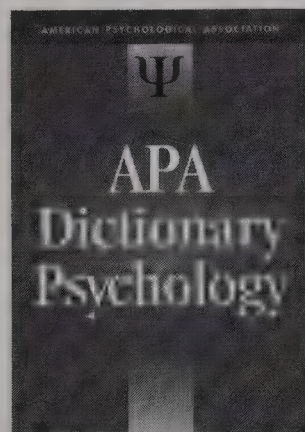
A handy abridgment of the award-winning guide to the language of the field from the world's largest professional association of psychologists, with many updates and revised definitions

- 10,000 entries offering clear and authoritative definitions—including many revised and updated definitions from the parent dictionary (*APA Dictionary of Psychology*)
- Balanced coverage across core areas of psychology—including cognitive, personality, social, developmental, health, clinical, experimental, neuropsychology, research methodology, and many others
- Thousands of incisive cross-references to deepen the user's understanding of related topics
- A "Quick Guide to Use" that explains stylistic and formal features at a glance
- Appendixes listing major figures in the history of psychology and psychological therapies and approaches
- Use as a standalone reference or as a portable alternative to the *APA Dictionary of Psychology*

2009 | 592 pages | Hardcover | List: \$39.95 | APA Member/Affiliate: \$39.95

ISBN 978-1-4338-0391-8 | Item # 4311009

**Offer expires October 31, 2008.**



**RUSA Outstanding Reference Source**

**NYLA Best of Reference**

**Choice Outstanding Academic Title!**

### APA Dictionary of Psychology

With over 25,000 terms and definitions, the *Dictionary* encompasses all areas of research and application, and includes coverage of concepts, processes, and therapies across all the major subdisciplines of psychology. Ten years in the making and edited by a distinguished editorial board of nearly 100 psychological scholars, researchers and practitioners, the *APA Dictionary of Psychology* is destined to become the **most authoritative** reference of its kind.

*"This is an excellent reference work that should be ordered by every institutional library and every psychology department to make it accessible to students, faculty, researchers, and practitioners in all social science and health-related arenas."*

— **PsycCRITIQUES**

*"For the world's largest and most well known association of psychologists to produce a dictionary gives the work an 'out of the box' authority and prestige that few works share... Summing up: Essential."* — **Choice**

*"Ten years in the making, the American Psychological Association's (APA) new dictionary was well worth the wait."* — **Library Journal**, starred review

2006 | 1,024 pages | Hardcover | List: \$59.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-59147-380-0 | Item # 4311009

AD0598



Volume 100  
Number 4

November 2008

Published quarterly  
by the  
American Psychological  
Association

ISSN 0022-0663

# Journal of Educational Psychology

Karen R. Harris, *Editor*

Eric M. Anderman, *Associate Editor*

Donna M. Kulikowich, *Associate Editor*

Gloria Miller, *Associate Editor*

Frank Pajares, *Associate Editor*

Jeffrey J. Walczyk, *Associate Editor*

**CURRENT YR/VOL**

**Marygrove College Library**  
8425 West McNichols Road  
Detroit, MI 48221

[www.apa.org/journals/edu](http://www.apa.org/journals/edu)

2000-2010  
**DECADE**  
*of* **BEHAVIOR**



## Editor

Karen R. Harris, *Vanderbilt University*

## Associate Editors

Eric M. Anderman, *Ohio State University*

Jonna M. Kulikowich, *Pennsylvania State University*

Gloria Miller, *University of Denver*

Frank Pajares, *Emory University*

Jeffrey J. Walczyk, *Louisiana Technical University*

## Chief Editorial Assistant

Brenna Hansen, *Vanderbilt University*

## Editorial Assistants

Karrie Godwin, *University of Denver*

Diana Griffith-Ross, *Louisiana Technical University*

Jason Chen, *Emory University*

Nicholas D. Warcholak, *Pennsylvania State University, University Park Campus*

## Advisory Editors

Patricia Alexander, *University of Maryland, College Park*

Ellen R. Altermatt, *Hanover College*

Lynley H. Anderman, *Ohio State University*

Robert Atkinson, *Arizona State University*

Carole Beal, *Information Sciences Institute at the University of Southern California*

Hefer Bembenuddy, *Queens College*

David A. Bergin, *University of Missouri—Columbia*

Benita A. Blachman, *Syracuse University*

Mimi Bong, *Ewha Womans University, Seoul, Korea*

Jere Brophy, *Michigan State University*

Scott W. Brown, *University of Connecticut*

Adriana G. Bus, *Leiden University, Leiden, the Netherlands*

Robert Calfee, *University of California, Riverside*

Joanne F. Carlisle, *University of Michigan*

Martha Carr, *University of Georgia*

Jerrell C. Cassady, *Ball State University*

Clark Chinn, *Rutgers University*

Namok Choi, *University of Louisville*

Donald L. Compton, *Vanderbilt University*

Alice J. Corkill, *University of Nevada, Las Vegas*

H. Michael Crowson, *University of Oklahoma*

Anne E. Cunningham, *University of California, Berkeley*

Teresa K. DeBacker, *University of Oklahoma*

Amanda M. Durik, *Northern Illinois University*

Pamela Beard El-Dinary, *Educational Consultant*

Dorothy L. Espelage, *University of Illinois at Urbana—Champaign*

Jill Fitzgerald, *University of North Carolina at Chapel Hill*

Douglas Fuchs, *Vanderbilt University*

Lynn S. Fuchs, *Vanderbilt University*

David C. Geary, *University of Missouri*

Alexandra Gottardo, *Wilfrid Laurier University, Waterloo, Ontario, Canada*

Steve Graham, *Vanderbilt University*

Barbara A. Greene, *University of Oklahoma*

Charles R. Greenwood, *University of Kansas*

John Guthrie, *University of Maryland, College Park*

Douglas J. Hacker, *University of Utah*

Vernon C. Hall, *Syracuse University*

Jenefer Husman, *Arizona State University*

Michael L. Kamil, *Stanford University*

Avi Kaplan, *Ben Gurion University of the Negev, Beer Sheva, Israel*

Robert M. Klassen, *University of Alberta, Edmonton, Alberta, Canada*

Beth Kurtz-Costes, *University of North Carolina at Chapel Hill*

Dan Lapsley, *University of Notre Dame*

Steve Lehman, *Utah State University*

Willy Lens, *University of Leuven, Leuven, Belgium*

Joel R. Levin, *University of Arizona*

Elizabeth A. Linnenbrink, *Duke University*

Mary Lundeborg, *Michigan State University*

Charles MacArthur, *University of Delaware*

Linda H. Mason, *Pennsylvania State University, University Park Campus*

Richard E. Mayer, *University of California, Santa Barbara*

Catherine McBride-Chang, *Chinese University of Hong Kong, Shatin, Hong Kong, China*

Valentina McInemey, *University of Western Sydney*

Debra K. Meyer, *Elmhurst College*

Michael Middleton, *University of New Hampshire*

Lisa M. Soederberg Miller, *University of California, Davis*

Raymond B. Miller, *University of Oklahoma*

Jens Möller, *University of Kiel, Kiel, Germany*

Tamera B. Murdock, *University of Missouri—Kansas City*

Karen P. Murphy, *Pennsylvania State University, University Park Campus*

Darcia Narvaez, *University of Notre Dame*

Markku Niemivirta, *University of Helsinki, Helsinki, Finland*

Jane Oakhill, *University of Sussex, Falmer, Brighton, United Kingdom*

Rollanda E. O'Connor, *University of California, Riverside*

Richard Olson, *University of Colorado*

Helen Patrick, *Purdue University*

Nancy Perry, *University of British Columbia, Vancouver, British Columbia, Canada*

Gary Phye, *Iowa State University*

Jan L. Plass, *New York University*

Robert Reid, *University of Nebraska—Lincoln*

Robert Renaud, *University of Manitoba, Winnipeg, Manitoba, Canada*

Alison M. Ryan, *University of Illinois at Urbana—Champaign*

Hollis S. Scarborough, *Haskins Laboratories, New Haven, Connecticut*

Christopher Schatschneider, *Florida State University*

Wolfgang Schneider, *Universität Würzburg, Würzburg, Germany*

Marlene Schommer-Aikins, *Wichita State University*

Gregory Schraw, *University of Nevada, Las Vegas*

Einar M. Skaalvik, *Norwegian University of Science and Technology, Trondheim, Norway*

Susan Sonnenschein, *University of Maryland, Baltimore County*

Laura M. Stapleton, *University of Maryland, Baltimore County*

Joseph Stevens, *University of Oregon*

H. Lee Swanson, *University of California, Riverside*

John Sweller, *University of New South Wales, Sydney, New South Wales, Australia*

Sonya Symons, *Acadia University, Wolfville, Nova Scotia, Canada*

Keith Thiede, *University of Illinois at Chicago*

Theresa A. Thorkildsen, *University of Illinois at Chicago*

Tim Urdan, *Santa Clara University*

Ellen Usher, *University of Kentucky*

Giovanni Valiante, *Rollins College*

Sharon Vaughn, *University of Texas at Austin*

Regina Vollmeyer, *University of Frankfurt, Frankfurt, Germany*

Charles A. Weaver III, *Baylor University*

Kathryn R. Wentzel, *University of Maryland, College Park*

Allan Wigfield, *University of Maryland, College Park*

Joanna P. Williams, *Teachers College, Columbia University*

Christopher A. Wolters, *University of Houston*

Moshe Zeidner, *University of Haifa, Haifa, Israel*

Barry J. Zimmerman, *Graduate Center, City University of New York*

The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

**Change of Address:** Send change of address notice and a recent mailing label to the attention of Subscriptions Department, APA, 30 days prior to the actual change of address. APA will not replace undelivered copies resulting from address changes; journals will be forwarded only if subscribers notify the local post office in writing that they will guarantee periodicals forwarding postage.

**Single Issues, Back Issues, and Back Volumes:** For information regarding single issues, back issues, or back volumes, write to Order Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

**Microform Editions:** For information regarding microform editions, write to University Microfilms, Ann Arbor, MI 48106.

**Manuscripts:** Effective in January 2008, the Incoming Editor is receiving all new submissions to the journal. Submissions that are accepted will be published beginning in the 2009 volume. Submit manuscripts electronically via the Manuscript Submission Portal at <http://www.apa.org/journals/edu/submission.html> according to the Instructions to Authors (see the table of contents). The opinions and statements published are the responsibility of the authors, and such opinions and statements do not necessarily represent the policies of APA or the views of the Editor.

**Copyright and Permission:** Those who wish to reuse APA-copyrighted material in a non-APA publication must secure from APA and the author of the material written permission to reproduce a journal article in full or journal text of more than 800 cumulative words or more than 3 tables or figures. APA normally grants permission contingent on permission of the author, inclusion of the APA copyright notice on the first page of reproduced material, and payment of a fee of \$25 per page. Permission from APA and fees are waived for those who wish to reproduce a single table or figure from a journal for use in a print product, provided the author's permission is obtained and full credit is given to APA as copyright holder and to the author through a complete citation. (Requesters requiring written permission for commercial use of a single table or figure will be assessed a \$25 service fee.) Permission and fees are waived for authors who wish to reproduce their own material for personal use; fees only are waived for authors who wish to use more than a single table or figure of their own material commercially (but for use in edited books, fees are waived for the author only if serving as the book editor). Permission and fees are waived for the photocopying of isolated journal articles for nonprofit classroom or library reserve use by instructors and educational institutions. A permission fee may be charged to the requester if students are charged for the material, multiple articles are copied, or large-scale copying is involved (e.g., for course packs). Access services may use unedited abstracts without the permission of APA or the author. Libraries are permitted to photocopy beyond the limits of the U.S. copyright law: (1) post-1977 articles, provided the per-copy fee in the code for this journal (0022-0663/08/\$12.00) is paid through the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923; (2) pre-1978 articles, provided that the per-copy fee stated in the Publishers' Fee List is paid through the Copyright Clearance Center. Address requests for reprint permission to Permissions Office, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

**Electronic Access:** APA members who subscribe to this journal have automatic access to all issues of the journal in the PsycARTICLES® full-text database. See <http://my.apa.org/access.html>.

**Reprints:** Authors may order reprints of their articles from the printer when they receive proofs.

**APA Journal Staff:** Susan J. A. Harris, *Senior Director, Journals Program*; Skip Maier, *Director, Journal Services*; Paige W. Jackson, *Director, Editorial Services*; Clark Munsell, *Account Manager*; Jodi Ashcraft, *Director, Advertising Sales and Exhibits*.

The **Journal of Educational Psychology** (ISSN 0022-0663) is published quarterly (February, May, August, November) in one volume per year by the American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Subscriptions are available on a calendar year basis only (January through December). The 2009 rates follow: *Nonmember Individual*: \$159 Domestic, \$184 Foreign, \$194 Air Mail. *Institutional*: \$490 Domestic, \$532 Foreign, \$545 Air Mail. *APA Member*: \$76. Write to Subscriptions Department, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242. Printed in the U.S.A. Periodicals postage paid at Washington, DC, and at additional mailing offices. POSTMASTER: Send address changes to the *Journal of Educational Psychology*, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.

# Educational Psychology

November 2008

Volume 100  
Number 4

[www.apa.org/journals/edu](http://www.apa.org/journals/edu)

## Articles

Copyright © 2008  
by the  
American  
Psychological  
Association

- 727 Effect of Retention in First Grade on Children's Achievement Trajectories Over 4 Years: A Piecewise Growth Analysis Using Propensity Score Matching  
*Wei Wu, Stephen G. West, and Jan N. Hughes*
- 741 Classroom Age Composition and Developmental Change in 70 Urban Preschool Classrooms  
*Arlen C. Moller, Emma Forbes-Jones, and A. Dirk Hightower*
- 754 Peer Victimization and Academic Achievement in a Multiethnic Sample: The Role of Perceived Academic Self-Efficacy  
*Jochem Thijs and Maykel Verkuyten*
- 765 Engagement and Disaffection in the Classroom: Part of a Larger Motivational Dynamic?  
*Ellen Skinner, Carrie Furrer, Gwen Marchand, and Thomas Kindermann*
- 782 Children's Early Interest-Based Activities in the Home and Subsequent Information Contributions and Pursuits in Kindergarten  
*Carin Neitzel, Joyce M. Alexander, and Kathy E. Johnson*
- 798 Supporting Students' Motivation, Engagement, and Learning During an Uninteresting Activity  
*Hyungshim Jang*
- 812 College Seniors' Theory of Their Academic Motivation  
*Shawn Van Etten, Michael Pressley, Dennis M. McInerney, and Arief Darmanegara Liem*
- 829 Dynamic Assessment of Algebraic Learning in Predicting Third Graders' Development of Mathematical Problem Solving  
*Lynn S. Fuchs, Donald L. Compton, Douglas Fuchs, Kurstin N. Hollenbeck, Caitlin F. Craddock, and Carol L. Hamlett*
- 851 Catching Up or Falling Behind? Initial English Proficiency, Concentrated Poverty, and the Reading Growth of Language Minority Learners in the United States  
*Michael J. Kieffer*
- 869 From Reading to Spelling and Spelling to Reading: Transfer Goes Both Ways  
*Nicole J. Conrad*
- 879 A Meta-Analysis of Single Subject Design Writing Intervention Research  
*Leslie Ann Rogers and Steve Graham*
- 907 Primary Grade Writing Instruction: A National Survey  
*Laura Cutler and Steve Graham*
- 920 Epistemological Beliefs' Contributions to Study Strategies of Asian Americans and European Americans  
*Marlene Schommer-Aikins and Marilyn Easter*

(Contents continue)



- 930 Heuristics and Biases as Measures of Critical Thinking: Associations with Cognitive Ability and Thinking Dispositions  
*Richard F. West, Maggie E. Toplak, and Keith E. Stanovich*
- 942 Identifying Patterns of Appraising Tests in First-Year College Students: Implications for Anxiety and Emotion Regulation During Test Taking  
*Heather A. Davis, Christine DiStefano, and Paul A. Schutz*
- 961 Confidence and Cognitive Test Performance  
*Lazar Stankov and Jihyun Lee*
- 977 The Role of Passion for Teaching in Intrapersonal and Interpersonal Outcomes  
*Noémie Carbonneau, Robert J. Vallerand, Claude Fernet, and Frédéric Guay*
- 988 Athletic Classmates, Physical Self-Concept, and Free-Time Physical Activity: A Longitudinal Study of Frame of Reference Effects  
*Ulrich Trautwein, Erin Gerlach, and Oliver Lüdtke*

## Other

- 1002 Acknowledgment of Ad Hoc Reviewers
- 987 American Psychological Association Subscription Claims Information
- 850 Call for Nominations
- ix Instructions to Authors
- 941 Low Publication Prices for APA Members and Affiliates
- 929 Members of Underrepresented Groups: Reviewers for Journal Manuscripts Wanted
- 868 New Editors Appointed, 2010–2015
- ii Subscription Order Form

## ORDER FORM

Start my 2009 subscription to the *Journal of Educational Psychology* ISSN: 0022-0663

\_\_\_ \$76.00 APA MEMBER/AFFILIATE \_\_\_\_\_

\_\_\_ \$159.00 INDIVIDUAL NONMEMBER \_\_\_\_\_

\_\_\_ \$490.00 INSTITUTION \_\_\_\_\_

*In DC add 5.75% / In MD add 6% sales tax* \_\_\_\_\_

**TOTAL AMOUNT DUE** \$ \_\_\_\_\_

**Subscription orders must be prepaid.** Subscriptions are on a calendar year basis only. Allow 4-6 weeks for delivery of the first issue. Call for international subscription rates.



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION

### SEND THIS ORDER FORM TO

American Psychological Association  
Subscriptions  
750 First Street, NE  
Washington, DC 20002-4242

Call **800-374-2721** or 202-336-5600  
Fax **202-336-5568** :TDD/TTY **202-336-6123**  
For subscription information,  
e-mail: [subscriptions@apa.org](mailto:subscriptions@apa.org)

☐ **Check enclosed** (make payable to APA)

**Charge my:** ☐ Visa ☐ MasterCard ☐ American Express

Cardholder Name \_\_\_\_\_

Card No. \_\_\_\_\_ Exp. Date \_\_\_\_\_

\_\_\_\_\_  
Signature (Required for Charge)

### Billing Address

Street \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

Daytime Phone \_\_\_\_\_

E-mail \_\_\_\_\_

### Mail To

Name \_\_\_\_\_

Address \_\_\_\_\_

City \_\_\_\_\_ State \_\_\_\_\_ Zip \_\_\_\_\_

APA Member # \_\_\_\_\_

EDUA09

# Effect of Retention in First Grade on Children's Achievement Trajectories Over 4 Years: A Piecewise Growth Analysis Using Propensity Score Matching

Wei Wu and Stephen G. West  
Arizona State University

Jan N. Hughes  
Texas A&M University

The authors investigated the relatively short-term and longer term effects of grade retention in 1st grade on the growth of mathematics and reading achievement over 4 years. The authors initially identified a large multiethnic sample ( $n = 784$ ) of children who were below the median in literacy at school entrance. From this sample, the authors closely matched 1 retained with 1 promoted child ( $n = 97$  pairs) on the basis of propensity scores constructed from 72 background variables and compared growth of retained and promoted children using Rasch-modeled W scores and grade standard scores, which facilitate age-based and grade-based comparisons, respectively. When using W scores, retained children experienced a slower increase in both mathematics and reading achievement in the short term but a faster increase in reading achievement in the longer term than did the promoted children. When using grade standard scores, retained children experienced a faster increase in the short term but a faster decrease in the longer term in both mathematics and reading achievement than did promoted children. Some of the retention effects were moderated by limited English language proficiency, home-school relationship, and children's externalizing problems.

**Keywords:** grade retention, growth curve model, propensity score, optimal matching, achievement

Since the mid-1990s, “ending social promotion” has become a central component of the standards-based reform movement that emphasizes setting content-based standards for students and holding both schools and students accountable for meeting them. In his 1998 and 1999 State of the Union addresses, President Clinton urged an end to social promotion and stated that scores on standardized tests should be the basis for promotion. The No Child Left Behind federal legislation passed in 2001, championed by President George W. Bush, requires that assessments aligned with state standards be used to measure the achievement of all children at each grade level. States increasingly use performance on these tests to evaluate schools and to reach decisions about promotion of children to the next grade.

Texas, the location of the present study, has been a leader in the standards-based reform movement. In 1990, the Texas legislature enacted a policy requiring all schools to test students in Grades 3 and higher with a test aligned with statewide curricular standards. Schools are evaluated on the basis of performance on this test, currently the Texas Assessment of Knowledge and Skills; test results are widely publicized; and rewards and sanctions are allo-

cated to schools on the basis of performance standards. In 1999, Texas enacted legislation that took effect in the 2002–2003 school year that requires students in Grades 3, 5, and 8 to pass the Texas Assessment of Knowledge and Skills in specified content areas to be promoted to the next grade, unless certain exceptional circumstances apply (e.g., documented learning disability). These policies were initiated by then-Governor George W. Bush and served as the model for the No Child Left Behind federal legislation.

Although the effectiveness of these educational accountability policies is hotly disputed (Evans, Baugh, & Sheffer, 2005; Hong & Raudenbush, 2005; Sipple, Killeen, & Monk, 2004; Warren & Edwards, 2005), there is little doubt that one consequence of the use of high-stakes testing is a decrease in the frequency of social promotion (defined as promoting to the next grade children who have not mastered curriculum content at their current grade) and an increase in grade retention (Gootman, 2005; Roderick, Bryk, Jacob, Easton, & Allensworth, 1999; Roderick & Nagaoka, 2005). In Texas, retention rates in Grades K–5 increased from 1994–1995 to 2003–2004, with the retention rate for Grade 3, the first promotional gate grade, up 100% over this time period (Texas Education Agency, 2005). The accountability movement likely affects retention policies even when performance on the curriculum-aligned test is not the sole or primary criterion for grade retention decisions. In Texas in 2003–2004, retention in first grade, the focal grade in this study, was 6.4%, compared with 5.8% in 1994–1995 (Texas Education Agency, 2005).

---

Wei Wu and Stephen G. West, Department of Psychology, Arizona State University; Jan N. Hughes, Department of Educational Psychology, Texas A&M University.

This research was supported in part by National Institute of Child Health and Development Grant 5 R01 HD39367-02 to Jan N. Hughes.

Correspondence concerning this article should be addressed to Wei Wu, who is now at the Department of Psychology, University of Kansas, Lawrence, KS 66045-7556, or Jan N. Hughes, Department of Educational Psychology, Texas A&M University, College Station, TX 77843-4225. E-mail: wwei@ku.edu or jhughes@tamu.edu

## Previous Research on Retention Effects

When a student has failed to demonstrate grade-level competencies, one option is to retain the student in hopes that another



year of maturity and exposure to the curriculum of the repeated grade will prepare the child to meet the academic and social demands of the next grade, thereby increasing the probability of academic and social success in future grades. Despite the intuitive appeal of the argument for grade retention as an intervention for poor achievement, the weight of available empirical evidence of varying methodological quality collected over 50 years suggests that grade retention either bestows no benefits on the retained student or has a negative impact on achievement and on social and emotional adjustment, self-confidence, and attachment to school (Dennebaum & Kulberg, 1994; Hong & Raudenbush, 2005; McCoy & Reynolds, 1999; Miesels & Liaw, 1993; Pagani, Tremblay, Vitaro, Boulerice, & McDuff, 2001; Reynolds & Bezruczko, 1993). Meta-analytic investigations of the published literature report overall negative effects of grade retention (Holmes, 1989; Holmes & Matthews, 1984; Jimerson, 2001). Grade retention has also been associated with a substantial increase in school withdrawal before high school completion (Jimerson, 1999; Roderick, 1994), even when retained students' academic performance is similar to that of comparably low-achieving promoted peers (Alexander, Entwisle, & Dauber, 2003).

Because children are not randomly assigned to the retention intervention, some critics (e.g., Hong & Raudenbush, 2005; Reynolds, 1992) have argued that many of the available studies that compare the postretention performance of retained children with that of nonselected promoted children tell us little about the impact of grade retention. Since the 1980s, the majority of studies on the effect of grade retention have compared the performance of retained students with that of low-achieving promoted peers. These studies have used two types of comparisons: (a) same-grade comparisons and (b) same-age comparisons. Same-grade comparisons evaluate retained and promoted students when they are in the same school grade. Thus, the performance of retained children is compared with promoted students who are, on average, a year younger. Alternatively, a proxy for this direct same-grade comparison can be used in which retained and promoted students are compared with extensive norms based on children in the same school grade. In contrast, same-age comparisons compare retained and nonretained children directly on measures of performance. Thus, the performance of retained children is compared with that of promoted students who are the same age but who will be one grade ahead of the target children. Both types of comparisons are informative, but they clearly answer different questions.

Results of these studies have been mixed, with some studies finding positive effects for achievement (Alexander, Entwisle, & Dauber, 1994; Mantzicopoulos & Morrison, 1992) and others reporting no effects or negative effects (Miesels & Liaw, 1993; Pianta, Tietbohl, & Bennett, 1997; Reynolds, 1992). When positive effects are reported, they are typically short-term effects that diminish within 2 or more years after the repeat year (Pierson & Connell, 1992). Studies using same-age comparisons compare the outcomes of retained children with those of children deemed at risk for promotion who were, nevertheless, promoted to the next grade. Results of these studies are more consistent in documenting negative effects of grade retention (Dennebaum & Kulberg, 1994; Hong & Raudenbush, 2005; Jimerson, 1999).

Despite the large number of published studies investigating the effects of grade retention, methodological limitations of these studies provide a weak basis for reaching conclusions about the

impact of grade retention. The basic problem is the challenge of making causal inferences in the absence of a randomized experimental design (West & Thoenmes, 2008). It is neither feasible nor ethically appropriate to randomly assign students to the "treatments" of retention and promotion. A large number of variables at the child, family, school, and district level are associated with the treatment selection (i.e., retention versus promotion) and with the measured outcomes (Reynolds, 1992; Willson & Hughes, 2006). Because of this potential selection bias, a finding that retained children experience more negative outcomes at some future point in their academic careers, compared with all promoted children, may often tell us little about the effect of retention on these outcomes. Whereas some researchers have made no attempt to control for preretention differences, in recent years researchers have attempted to deal with this confounder in one or both of two ways.

In the first approach, researchers compare the performance of retained students with that of a comparison group of students who were deemed as being at risk for being retained but who were promoted. The typical approach is to select students in the same grade as the retained students who scored below a specific score (e.g., below the 25th or 50th percentile) on a measure of achievement or cognitive ability during that year, but who were promoted to the next grade the subsequent year. Some studies using this approach document equivalent performance for retained and promoted students on preretention measures. Unfortunately, there is no guarantee with this approach that the promoted and retained groups are fully equivalent on the measured variable. In addition, potential differences between promoted and retained children on other important variables known to be related to school performance (e.g., child hyperactivity, parental education level, and peer acceptance) may exist. These sources of nonequivalence from selection bias confound the interpretation of the effects of retention on outcomes (Reichardt, 2006).

In the second approach, statistical adjustments are made for a limited number of preretention variables, using analysis of covariance or, equivalently, multiple regression (Cohen, Cohen, West, & Aiken, 2003; Huitema, 1980). These statistical adjustment procedures make three key assumptions. First, the limited number of covariates that can be included in the model adequately capture the important preexisting differences between the retained and promoted groups. Second, the relationship (typically linear) between each covariate and the outcome is correctly specified. Third, the regression lines for the retained and nonretained groups are parallel. Researchers have rarely reported checking these critical assumptions, nor have they used alternative procedures that relax one or more of them (see, e.g., Little, Hyonggin, Johanns, & Giordani, 2000). In addition, when groups are disparate at pretest, the application of statistical adjustment procedures often assumes that the effect of the treatment (retention) can be extrapolated beyond the region in which the baseline data for the two groups actually overlap, often a very risky procedure (Shadish, Cook, & Campbell, 2002; Shadish, Luellen, & Clark, 2006).

During the past 2 decades, a new method of equating groups, propensity score analysis (McCaffrey, Ridgeway, & Morral, 2004; Rosenbaum, 2002; Rosenbaum & Rubin, 1983; Shadish et al., 2006; West & Thoenmes, 2008), has been extensively developed in statistics. Propensity score analysis offers important benefits over previous equating approaches, and it has begun to be used in



applied research on important social issues in which randomization is not possible. In this approach, participants are measured on a wide variety of baseline measures believed on the basis of prior substantive theory and research to be related to treatment selection, the outcome variable, or ideally both. To the extent that researchers have identified the important variables related to both treatment selection and outcome, it is possible to remove baseline differences between treatment groups and achieve unbiased estimates of treatment effects. A statistical model is used to estimate a single propensity score for each participant, the predicted probability that the participant will be in the treatment (retention) condition. Typically, logistic regression using all measured baseline variables as predictors is used for this step, although other, more complicated statistical models are occasionally needed.<sup>1</sup> The propensity score is then used to equate the treatment and control groups using matching, blocking (creating strata), or analysis of covariance. Both statistical theory (Rosenbaum, 2002; Rosenbaum & Rubin, 1983) and empirical research (Shadish & Clark, 2006) have shown that the use of propensity scores substantially reduces or eliminates selection bias when properly used. Careful use of propensity scores achieves approximate balance on the baseline levels of each of a comprehensive set of measured variables. Hong and Raudenbush (2005) recently introduced propensity score methods as a device to equate retained and promoted children in studies of grade retention.

### Study Purpose

The purpose of the current study was to combine three methods that help maximize the internal validity of nonrandomized studies to examine the effect of retention in first grade on children's growth in reading and math over a 4-year period, beginning with the child's 1st year in first grade. First, before any of the children were retained, we selected and comprehensively assessed a sample of children at risk for retention on the basis of their scores below the 50th percentile on school district tests of reading. Second, we used propensity score matching, which corrects for selection bias, using an extensive, carefully chosen set of variables measured at baseline. Third, we used growth curve modeling that permits estimation of each child's trajectory of growth in mathematics and reading achievement. With four waves of data, we can examine the effects of retention on both short-term and longer term growth in math and reading achievement. Specifically, we used piecewise linear trajectory models (Singer & Willett, 2003) to investigate both immediate and longer term effects. We expected the slope for retention effects in the interval from Year 1 through Year 2 would differ from that in the interval from Year 2 through Year 4 for the retained children. Otherwise stated, we could compare the differences in the growth of the retained and promoted children in the short term, when retained children are repeating the first-grade curriculum, and in subsequent years, when retained and promoted children are exposed to novel curriculum at new grade levels.

We analyzed achievement results using both grade-level standard scores and W scores (a Rasch-type measure of ability) from the Woodcock-Johnson III (WJ-III), a well-researched, nationally standardized measure of reading and math achievement. Grade-level standard scores compare students with grade-level norms based on the student's current grade. W scores are a measure of actual growth in math and reading ability; comparisons between

retained and promoted children using W scores are comparable to age norms, in that retained children's rate of growth in the underlying, latent construct of math or reading is directly compared with that of promoted children for the same time interval. We expected that results would differ on the basis of whether grade standard scores or W scores were analyzed. Specifically, using grade standard scores, we expected retention would favor retained children during the short term but that the benefit would begin to dissipate when retained students were promoted to the second grade. Using W scores, we expected retention to favor promoted children in the short term, when promoted students are exposed to a new curriculum. We also hypothesized that the slopes representing the longer term growth for retained and promoted students would either be comparable (parallel) after the repeat year or, if the promise of retention were fulfilled, higher for retained students because of their stronger foundation in the content covered in the repeat grade.

Finally, we investigated whether several factors at various levels of analysis (i.e., child, family, and classroom) moderate the effects of grade retention. Consistent with developmental systems theory (Lerner, 1989), we expected that the impact of retention on a child's achievement trajectory would differ on the basis of the interplay of the retention "treatment" and factors at multiple levels of analysis, including factors within and outside the child (Cicchetti & Posner, 2005; Sameroff, 1975, 1989). Previous investigations of moderator variables have been restricted to distal demographic variables such as gender, race, or family socioeconomic status, with inconsistent findings (Pagani et al., 2001; Reynolds, 1992). We investigated more proximal child and classroom variables that were expected to moderate the effects of retention. Specifically, we examined children's behavioral regulation, as indexed by teacher, peer, and parent ratings of externalizing behaviors, child personality resilience, and teacher-student and home-school relationship quality. We expected that the relative benefit of promotion versus retention on children's achievement trajectories would be stronger for children with better behavioral regulation, greater personality resilience (agreeable, conscientious, and persistent), and more supportive teacher-student and home-school relationships. We reasoned that such child assets would enable children who are promoted, relative to their propensity-matched retained children, to successfully meet the greater maturity and academic demands of higher grades. We also tested age as a moderator because parents and educators are more likely to retain younger children for age, relative to similar low-achieving students (Mantzicopoulos, 2003; Reynolds, 1992; Willson & Hughes, 2006). Thus, it is important to evaluate the wisdom of using age as a selection factor for the retention treatment.

This study extends a previous study (Wu, West, & Hughes, 2008) with this sample. Because only three waves, or years, of data were available in the Wu et al. (2008) study, short-term and longer term change could not be investigated. That study found negative effects for retention on math W scores but no effects for reading W scores. Finally, the short-term beneficial effects of promotion were

<sup>1</sup> Logistic regression is the most commonly used method of estimating propensity scores, although other approaches such as regression tree methods can also be used (see McCaffrey et al., 2004, and West & Thommes, 2008, for a discussion of the major estimation methods).



stronger for children with good behavioral regulation and for children who were not classified as Limited English Proficient.

The current study analyzes both grade standard scores and W scores, whereas the previous study analyzed only W scores. The analysis of both W scores and grade standard scores permits us to answer two different questions. Specifically, because W scores are an interval-level measure of reading and math skills, their use permits a determination of the effect of grade retention on children's growth in achievement. Because grade standard scores compare a student's performance to grade-level norms for the same grade as that in which the student is enrolled, it permits a determination of the effect of grade retention on children's performance relative to the student's current grade placement. Both comparisons have merit (Alexander et al., 2003). By modeling both short-term and longer term effects and by analyzing both children's performance relative to grademates and children's actual growth in reading and math, we are able to render a more comprehensive picture of the impact of grade retention than has been possible in previous studies. Research by Miles and Stipek (2006) and Skinner, Zimmer-Gemback, and Connell (1998) has suggested that children's relative (rank order) level of achievement becomes relatively stable after Grade 3. Although their research did not investigate children's achievement trajectories, which is the focus of the present study, it does suggest that understanding the impact of early grade retention may potentially be important in the understanding of children's long-term achievement outcomes.

## Method

### Participants

Participants were drawn from a larger sample of children participating in a longitudinal study examining the impact of grade retention on academic achievement. Participants were recruited from three school districts in Texas (one urban and two small cities) across two sequential cohorts in first grade during the fall of 2001 and 2002. Children were eligible to participate in the longitudinal study if they scored below the median score on a state-approved district-administered measure of literacy, spoke either English or Spanish, were not receiving special education services, and had not previously been retained in first grade. School records identified 1,374 children as eligible to participate. Because teachers distributed consent forms to parents via children's weekly folders, the exact number of parents who received the consent forms cannot be determined. Incentives in the form of small gifts to children and the opportunity to win a larger prize in a random drawing were instrumental in obtaining 1,200 returned consent forms, of which 784 parents (65%) provided consent and 416 declined.

Analyses of a broad array of archival variables including performance on the district-administered test of literacy (standardized within district because of differences in the test used), age, gender, ethnicity, eligibility for free or reduced-price lunch, bilingual class placement, cohort, and school context variables (i.e., percentage ethnic or racial minority and percentage economically disadvantaged) did not indicate any differences between children with and without consent. The resulting sample of 784 participants (52.6% male) closely resembles the population from which they were drawn on demographic and literacy variables relevant to students'

educational performance. The ethnic composition of the achieved sample ( $N = 784$ ) was 37% Hispanic (39% of whom were Spanish language dominant), 34% White Caucasian, 23% African American, and 6% other; 62% of the children qualified for free or reduced-price lunch. The mean full scale IQ based on the Universal Nonverbal Intelligence Test (Bracken & McCallum, 1998) for the sample was 92.91 ( $SD = 18.01$ ), and the mean reading achievement score was 96.40 ( $SD = 14.28$ ).

Participants for the growth curve analyses were those 196 children (58% male) who were successfully matched with respect to their propensity to be retained in first grade (see description of propensity-matching procedures below) and who had achievement test scores for at least one of the four assessment periods. The racial and ethnic composition of the sample was 33% Caucasian, 33% Hispanic, 31% African American, and 3% other. At entrance to first grade, children's mean age was 6.46 ( $SD = 0.33$ ) years. Of the 196 matched cases (98 pairs), the proportions of children who completed the WJ-III were 96%, 85%, 86%, and 85%, respectively, at each of the four measurement waves.

### Design Overview

Demographic information including child age, gender, race or ethnicity, eligibility for free or reduced-price lunch, and status as Limited English Proficient were obtained from school district records. Teacher, parent, and peer data were collected in the spring of Year 1, when all participants were in first grade. Teachers and parents received \$25.00 for completing and returning the questionnaires. Peers' perceptions of the level of externalizing behaviors were obtained via individual interviews conducted between February and May of Year 1. Beginning in Year 1, annual measures of math and reading achievement were individually administered at school for 4 years, with the constraint that at least 8 months separated each annual assessment.

### Measures

A complete list of the 72 baseline variables collected at measurement Wave 1 used in calculation of propensity scores is available from Jan N. Hughes. The baseline variables included demographic measures, cognitive and behavioral performance, social and emotional functioning, and classroom and school variables. The 72 variables were intended to be as comprehensive as possible, including variables that have been shown in prior research to be related to early retention versus promotion, early academic achievement, or ideally both. The measures of academic achievement served as the primary outcome measures for this study. Selected baseline measures were used to explore potential moderator variables that might affect the rate of growth of children in response to retention versus promotion. These measures and the achievement measures are described below.

**Academic achievement.** The WJ-III Tests of Achievement (Woodcock, McGrew, & Mather, 2001) are individually administered measures of academic achievement for individuals ages 2 to adulthood. We used the WJ-III Broad Reading (Letter-Word Identification, Reading Fluency, and Passage Comprehension subtests) and Broad Math (Calculations, Math Fluency, and Math Calculation Skills subtests) W scores and grade standard scores. The Reading and Math W scores are based on the Rasch measurement



model, which ensures interval-level measurement of change on a single dimension.<sup>2</sup> Grade standard scores compare children's performance to grade-level national norms and only approximate an equal-interval measurement scale. Extensive research has documented the reliability and construct validity of the WJ-III and its predecessors (Woodcock & Johnson, 1989; Woodcock et al., 2001).

The Bateria Woodcock-Muñoz: Pruebas de Aprovechamiento—Revisada (Bateria-R; Woodcock & Muñoz-Sandoval, 1996) is the comparable Spanish version of the Woodcock-Johnson Tests of Achievement—Revised (WJ-R; Woodcock & Johnson, 1989), the precursor to the WJ-III. If children or their parents spoke any Spanish, children were administered the Woodcock-Muñoz Language Test (Woodcock & Muñoz-Sandoval, 1993) to determine the child's language proficiency in English and Spanish. The test of achievement (WJ-III or the Bateria-R) was chosen to match the language in which the child had greater proficiency. The Woodcock Compuscore (Woodcock & Muñoz-Sandoval, 2001) program yields W scores for the Bateria-R that are comparable to W scores on the WJ-R. In the following, scores are referred to as WJ scores (W or grade standard scores), irrespective of which test the child took.

*Teacher and parent report of conduct problems and hyperactivity.* Teachers and parents completed the Strengths and Difficulties Questionnaire (Goodman, 1997), a brief (25-item) screening measure for psychopathology. Each item is rated on a scale ranging from 0 to 2 (i.e., *not true*, *somewhat true*, and *certainly true*). The Strengths and Difficulties Questionnaire yields five scales consisting of 5 items each. The Conduct Problems and the Hyperactivity scales assess externalizing behaviors. For our sample, coefficient alpha for Conduct Problems was .84 for teachers and .71 for parents. For Hyperactivity, coefficient alpha was .89 for teachers and .81 for parents. In a sample of children participating in this longitudinal study, teacher reports of Conduct Problems and of Hyperactivity were moderately to strongly correlated with both parent report (.47 and .30, for conduct problems and hyperactivity, respectively) and peer reports (.50 and .46, for conduct problems and hyperactivity, respectively; Hill & Hughes, 2007). Exploratory and confirmatory factor analyses support the construct validity of the teacher and parent versions of the Strengths and Difficulties Questionnaire (Dickey & Blumberg, 2004; Goodman, 2001; Hill & Hughes, 2007).

*Peer nomination of externalizing problems.* Peers' perceptions of classmates' hyperactivity and aggression were obtained following procedures widely recommended in the peer assessment literature (Cillessen & Bukowski, 2000). Scores of similar constructs obtained from similar peer nomination procedures have demonstrated good reliability and validity (Realmuto, August, Sieler, & Pessoa-Brandao, 1997). In individual interviews, children were presented a roster with the names of all classmates. The interviewer read each of the classmates' names and asked the child whether he or she knew each child. Then the interviewer asked the child to nominate as few or as many classmates as he or she wished who fit each descriptor. Of interest to this study are the aggression item ("Some kids start fights, say mean things, or hit others") and the hyperactivity item ("Some kids do strange things and make a lot of noise. They bother people who are trying to work"). Each class member received an aggression and hyperactivity score based on the number of nominations that child received. Socio-

metric scores were standardized within classrooms. Because the two scores were highly correlated ( $r = .74$ ), we computed a composite peer-rated externalizing score as the mean standardized score on the aggression and hyperactivity items. Written parent consent was obtained for each child who participated in the sociometric interview. However, all children in a classroom were eligible to be rated or nominated. Terry (1999) reported that reliable and valid sociometric data can be collected using the unlimited nomination approach when as few as 40% of children in a classroom participate. We followed Terry's guideline, computing sociometric scores only for those 602 (77%) children located in classrooms in which more than 40% of classmates participated in the sociometric assessment. The mean rate of classmate participation in the sociometric administrations was .65 (range = .40 to .95). Elementary school children's peer nomination scores derived from procedures similar to those used in this study have been found to be stable over periods from 6 weeks to 4 years and to be associated with concurrent and future behavior and adjustment (for review, see Hughes, 1990).

*Resilient personality.* A confirmatory factor analysis (Kwok, Hughes, & Luo, 2007) on a sample of 445 first-grade children participating in the current longitudinal study supported a second-order measurement model of resilient personality defined by three first-order factors: Agreeableness (nine items), Conscientiousness (eight items), and Ego Resiliency (seven items). Agreeableness and Conscientiousness items were taken from the scales of the same name of the Big Five Inventory (John & Srivastava, 1999). Sample Agreeableness items include "is helpful and unselfish with others," "likes to cooperate with others," and "is sometimes rude to others" (reverse scored). Sample Conscientiousness items include "does a thorough job," "is a reliable worker," and "tends to be disorganized" (reverse scored). Coefficient alpha for each scale was .94. Ego Resiliency items were derived from items on the California Child Q-Sort (Caspi, Block, Block, & Klopp, 1992). Sample items include "resourceful in initiating necessary activities" and "falls to pieces under stress" (reverse scored). Teachers responded to each of the items using a 5-point scale. Coefficient alpha for our sample was .85. We computed a score for resilient personality as the mean of the standardized score for each scale.

*Teacher-student relationship quality.* The 22-item Teacher Student Relationship Inventory (Hughes, Cavell, & Willson, 2001) is based on the Network of Relationships Inventory (Buhrmester & Furman, 1987). Teachers indicated on a 5-point Likert-type scale their level of support (16 items, coefficient  $\alpha = .94$ ) or conflict (6 items, coefficient  $\alpha = .92$ ) in their relationships with individual students. Because the Support and Conflict scales were negatively correlated ( $-.57$ ), we recoded the Conflict items in the opposite direction. The coefficient alpha of the 22 items after recoding the Conflict items was .95. A total relationship quality score was computed as the mean of the 22 item scores. In a longitudinal study of behaviorally at-risk elementary school students, the Teacher Student Relationship Inventory Support score predicted changes in behavioral adjustment and peer relationships (Meehan, Hughes, & Cavell, 2003). In the larger sample, the Teacher Student Relationship Inventory

<sup>2</sup> In contrast, measures of reliability within classical test theory do not ensure that a single dimension has been measured or that interval-level measurement has been achieved.



predicted cross-year changes in children's achievement (Hughes & Kwok, 2006).

*Home-school relationship.* The Teacher Report of Parent Involvement Scale (Wong & Hughes, 2006) consists of 20 items assessing teachers' perceptions of the parent-teacher alliance (e.g., "I respect this parent" and "Communication between us is difficult"; reverse scored) and of the frequency of parents' engagement of various involvement activities (e.g., volunteering at school and calling the teacher). Coefficient alpha for the current sample was .87.

### Propensity Score Estimation

Propensity scores, the predicted probability of being retained in first grade, were estimated for the 768 children for whom retention information was available, using 72 background variables collected at the initial testing, including child demographic variables and child, peer, teacher, and parent data covering the areas of academic aptitude (e.g., the Universal Nonverbal Intelligence Test), academic achievement (WJ-III or the Spanish-language Batería-R broad math and reading), personality (e.g., agreeableness and effortful control), behavioral and social adjustment, peer relations, and family adversity. Methods based on logistic regression (Rosenbaum, 2002; Rosenbaum & Rubin, 1983) were used to estimate propensity scores. The larger the propensity score, the larger the predicted probability that the child would be retained in the first grade. For the 768 cases, the propensity score ranged from .0003 to .989 with a mean of .215 ( $SD = .215$ ). The children who were subsequently promoted had substantially lower propensity scores ( $n = 603$ ,  $M = .114$ ,  $SD = .166$ ) than those who were subsequently retained ( $n = 165$ ,  $M = .583$ ,  $SD = .309$ ),  $t(190.6) = -18.769$ , Cohen's  $d = -1.890$ . Although not the primary criterion in evaluating the success of the propensity model (see Rosenbaum 2002; Shadish et al., 2006), the logistic regression equation led to relatively good prediction of the decision to retain or promote each child with a Nagelkerke pseudo- $R^2$  index of .552 (see Cohen et al., 2003, p. 503).

Despite identifying an at-risk sample of children who were below the median on literacy at entrance to first grade, substantial differences existed between the retained and the promoted groups. These results indicated that an adjustment procedure would be needed to equate the retained and the promoted groups. Following the recommendations of West and Thoenes (2008; see also Rosenbaum, 2002), we chose a procedure that produces optimal matches on propensity scores.

### Matching Procedure

We matched 1 retained child with 1 promoted child on the basis of their propensity scores using SAS 8.0 PROC ASSIGN (Ming & Rosenbaum, 2001). PROC ASSIGN matches retained children with promoted children so that the sum of distance between the propensity scores within each of the matched pairs was minimized for the whole sample. To avoid matching two children with propensity scores far away from each other, we imposed a caliper distance of .025, the maximum distance in propensity scores allowed for a match to take place. That is, any pair of retained and promoted children who differed in their propensity scores by more than .025 could not be matched with each other. We chose a very small caliper distance to obtain high-quality matching. Using this method, a total of 98 pairs (196 children) were successfully

matched. For the 98 matched pairs, the propensity score ranged from .003 to .918 ( $M = .367$ ,  $SD = .225$ ). The mean within-pair distance in propensity score was .007 ( $SD = .008$ ). Following the optimal matching process, the two groups were closely equated on their propensity scores: For the promoted group,  $N = 98$  ( $M = .366$ ,  $SD = .225$ ), and for the retained group,  $N = 98$  ( $M = .367$ ,  $SD = .226$ ),  $t(194) = -0.044$ , *ns*, Cohen's  $d = -0.006$ . This outcome contrasts sharply with the substantial mean differences in the propensity scores between the promoted and retained groups in the full sample ( $N_{\text{promoted}} = 603$ ;  $N_{\text{retained}} = 165$ ). Table 1 reports descriptive information for the 98 pairs.

### Data Analysis

To investigate the relatively short-term and longer term effects of retention on the growth rate of WJ scores, we split the time span into two pieces at the point at which the measurement wave was 2 (the 2nd year of the study). The first piece included the measurements on Waves 1 and 2, which covers the change in WJ scores from roughly 0.5 year before retention to 0.5 year after retention. The second piece included the measurements on Waves 2, 3 and 4, which covers the change in WJ scores from roughly 0.5 year after retention to 2.5 years after retention. Then we fit a linear growth curve on each piece for the four WJ scores separately using SAS 8.0 PROC Mixed. Such growth curve models are called two-piece linear growth curve models (Singer & Willett, 2003). Given the existence of missing data, we used full information maximum likelihood estimation to estimate the growth curve models. Full information maximum likelihood estimation uses all of the observations available for each case to compute the likelihood function (Enders & Bandalos, 2001). It provides unbiased estimates with minimal standard errors when data are missing at random (Schafer & Graham, 2002). Otherwise stated, full information maximum likelihood estimation provides estimates that are appropriately corrected for all measured variables included in the analysis.

The specification of the two-piece linear growth curve model for any one of the WJ scores is shown in Equations 1 to 3, which includes three levels. Level 1 (within individual, Equation 1) captures the two-piece linear growth trajectory for each individual over the 4 years of the study. At Level 1, two time variables ( $T1_{\text{rip}}$  and  $T2_{\text{rip}}$ ), corresponding to the two pieces, were created to predict the WJ scores. To simplify the presentation, two coding schemes were adopted to create  $T1_{\text{rip}}$  and  $T2_{\text{rip}}$  as follows.

Coding Scheme 1:

	Wave 1	Wave 2	Wave 3	Wave 4
T1	0	1	1	1
T2	0	0	1	2

Coding Scheme 2:

	Wave 1	Wave 2	Wave 3	Wave 4
T1	0	1	2	3
T2	0	0	1	2

Following guidelines in Cohen et al. (2003, chap. 8), we used two coding schemes in the study even though their results are algebraically related and produce identical results for overall model fit and predicted values at each measurement wave. The use of the two schemes permitted us to conduct a significance test of

Table 1  
Descriptive Statistics of Selected Time 1 Measures

Time 1 measure	Retained ( <i>n</i> = 98)				Promoted ( <i>n</i> = 98)				Effect size	
	<i>M</i>	<i>SD</i>	<i>n</i>	%	<i>M</i>	<i>SD</i>	<i>n</i>	%	<i>d</i>	<i>h</i>
WJ-III math W score	461.85	12.54			460.85	15.77			0.07	
WJ-III reading W score	423.13	26.19			423.52	21.79			-0.02	
WJ-III math grade score	98.92	13.76			100.27	15.36			-0.09	
WJ-III reading grade score	89.43	17.32			92.95	12.83			-0.23	
Age at eligibility determination	6.45	0.29			6.46	0.36			-0.03	
Parent-rated conduct problems	0.54	0.50			0.48	0.44			0.13	
Parent-rated hyperactivity	1.25	0.57			1.09	0.48			0.30	
Teacher-rated conduct problems	0.40	0.50			0.52	0.57			-0.22	
Teacher-rated hyperactivity	0.95	0.63			1.06	0.61			-0.18	
Peer-rated externalizing problems	0.30	0.99			0.25	0.97			0.05	
Teacher-student relationship quality	3.99	0.75			3.92	0.83			0.09	
Home-school relationship	3.45	0.53			3.28	0.46			0.34	
Resilient personality	-0.12	0.84			-0.29	0.87			0.20	
Economically disadvantaged			65	66			63	66		.00
Ethnicity										
African American			33	34			27	28		.06
Hispanic			30	31			34	35		.04
Caucasian			32	33			35	36		.03
Others			1	1			2	2		.10

Note. *d* is reported as a measure of the standardized effect size for the difference between two means and *h* is reported as the standardized effect size for the difference between two proportions (Cohen, 1977). For both measures, Cohen suggested 0.2, 0.5, and 0.8 represented small, medium, and large effects, respectively. WJ-III = Woodcock-Johnson III.

each of the parameters of interest and to present the simplest possible interpretation of the results. Coding Scheme 1 allowed us to test the slope separately for each piece and furthermore to test whether the slopes in each piece differ between retained and promoted groups. However, Coding Scheme 1 does not permit us to test the significance of the difference in the change in slope between the short and longer term. That is one reason why we used Coding Scheme 2, which provides both this significance test and a test of whether the pattern of change in slope differs between the retained and the promoted groups. We centered time at Wave 1 (the point at which both  $T1_{tip}$  and  $T2_{tip}$  were coded 0); thus, the intercept represents the student's predicted level of WJ scores on the first wave of measurement (initial status in first grade before retention or promotion).

Level 2 (between individual, Equation 2) captures the variation in individual intercepts and slopes across individuals. At Level 2, we used grade retention status after first grade (coded 1 = retained, 0 = promoted) to predict the individual intercepts, individual slopes for the short term (Slope 1), and individual slopes for the longer term (Slope 2). Because matching produces dependency in the data, Level 3 (between pairs, Equation 3) was added to take into account the within-pair correlation for individual intercepts (clustering). This procedure adjusts for biased estimates of the standard errors caused by the dependency, thus leading to more accurate significance tests for the parameter estimates. The effects of retention on Slope 1 (short-term effect,  $\lambda_{110}$ ) and Slope 2 (longer term effect,  $\lambda_{210}$ ) are our primary interest. The equations for the three-level model are presented below.

$$\text{Level 1: } Y_{tip} = \pi_{0ip} + \pi_{1ip}T1_{tip} + \pi_{2ip}T2_{tip} + e_{tip};$$

$$e_{tip} \sim N(0, \sigma^2). \quad (1)$$

$$\text{Level 2: } \pi_{0ip} = \beta_{00p} + \beta_{01p}RETENTION_{ip} + r_{0ip};$$

$$\pi_{1ip} = \beta_{10p} + \beta_{11p}RETENTION_{ip}; \text{ and}$$

$$\pi_{2ip} = \beta_{20p} + \beta_{21p}RETENTION_{ip}; r_{0ip} \sim N(0, \tau_{\pi 00}). \quad (2)$$

$$\text{Level 3: } \beta_{00p} = \gamma_{000} + u_{00p}; \beta_{10p} = \gamma_{100}; \beta_{20p} = \gamma_{200};$$

$$\beta_{01p} = \gamma_{010}; \beta_{11p} = \gamma_{110}; \beta_{21p} = \gamma_{210}; u_{00p} \sim N(0, \tau_{\beta 00}). \quad (3)$$

Here the subscript *t* indicates time point (Wave 1, 2, 3, or 4), *i* indicates individual, and *p* indicates pair.  $\gamma_{000}$  and  $\gamma_{100}$  represent, respectively, the grand mean of the intercept and Slope 1 for the promoted group. Under Coding Scheme 1,  $\gamma_{200}$  represents the grand mean of Slope 2 for the promoted group. Under Coding Scheme 2,  $\gamma_{200}$  represents the average slope difference between the two pieces (Slope 2 – Slope 1) for the promoted group.  $\gamma_{010}$  and  $\gamma_{110}$  represent the estimated effects of grade retention in first grade on the intercept and Slope 1, respectively. Under Coding Scheme 1,  $\gamma_{210}$  represents the effect of grade retention on Slope 2, whereas under Coding Scheme 2,  $\gamma_{210}$  represents the effect of retention on the slope difference between the two pieces.  $e_{tip}$  represents the Level 1 residual for the *i*th individual within the *p*th pair at Wave *t*, which was assumed to follow a normal distribution with mean ( $\mu$ ) = 0 and homogeneous variance ( $\sigma^2$ ) across 4 years.  $\sigma^2$  is the within-individual variance in WJ scores that cannot be accounted for by time.  $r_{0ip}$  represents Level 2 residual in intercept for the *i*th individual within the *p*th pair, which was assumed to follow a normal distribution with  $\mu = 0$  and variance =  $\tau_{\pi 00}$ .  $\tau_{\pi 00}$  is the between-individual variance in intercept that cannot be accounted for by grade retention.  $u_{00p}$  represents the deviation of the mean intercept of the *p*th pair from the grand mean



intercept for promoted children, which was also assumed to follow a normal distribution with  $\mu = 0$  and variance =  $\tau_{\beta 00}$ .  $\tau_{\beta 00}$  is the between-pair variance in intercepts. Given that only four measurement waves were available to estimate a complex model, we assumed that there is neither between-individual residual variance nor between-pair variance in Slope 1 and Slope 2.

Next, the potential moderator variables were added into Level 2 of the growth model one at a time as shown in Equation 4. The moderating effects of these variables were captured by the coefficients associated with their interaction with grade retention ( $\gamma_{130}$  for Slope 1 and  $\gamma_{230}$  for Slope 2 with Coding Scheme 1). The moderating effects are assumed to be constant across matched groups (see Equation 5). We only used Coding Scheme 1 in the moderator analyses for ease of presentation because we could interpret the moderating effects on Slope 1 and Slope 2 separately.

Level 2:  $\pi_{0ip} = \beta_{00p} + \beta_{01p}RETENTION_{ip}$   
 $+ \beta_{02p}MODERATOR_{ip}$   
 $+ \beta_{03p}MODERATOR_{ip} \times RETENTION_{ip}$   
 $+ r_{0ip};$   
 $\pi_{1ip} = \beta_{10p} + \beta_{11p}RETENTION_{ip} + \beta_{12p}MODERATOR_{ip}$   
 $+ \beta_{13p}MODERATOR_{ip} \times RETENTION_{ip};$  and  
 $\pi_{2ip} = \beta_{20p} + \beta_{21p}RETENTION_{ip} + \beta_{22p}MODERATOR_{ip}$   
 $+ \beta_{23p}MODERATOR_{ip} \times RETENTION_{ip}.$  (4)

Level 3:  $\beta_{00p} = \gamma_{000} + u_{00p}; \beta_{10p} = \gamma_{100}; \beta_{20p} = \gamma_{200};$   
 $\beta_{01p} = \gamma_{010}; \beta_{02p} = \gamma_{020}; \beta_{03p} = \gamma_{030};$   
 $\beta_{11p} = \gamma_{110}; \beta_{12p} = \gamma_{120}; \beta_{13p} = \gamma_{130};$   
 $\beta_{21p} = \gamma_{210}; \beta_{22p} = \gamma_{220}; \beta_{23p} = \gamma_{230}.$  (5)

Here  $\gamma_{000}$ ,  $\gamma_{100}$ , and  $\gamma_{200}$  represent the estimated mean intercept, Slope 1, and Slope 2 for the promoted group with 0 value on moderator, respectively;  $\gamma_{010}$ ,  $\gamma_{110}$ , and  $\gamma_{210}$  represent the main effects of grade retention on the intercept and slope, respectively;  $\gamma_{020}$ ,  $\gamma_{120}$ , and  $\gamma_{220}$  represent the main effects of moderator on the intercept, Slope 1, and Slope 2, respectively; and  $\gamma_{030}$ ,  $\gamma_{130}$ , and  $\gamma_{230}$  represent the interaction effects of moderator and retention on the intercept, Slope 1, and Slope 2, respectively. The interpreta-

tions of Level 1 and Level 2 residuals are similar with the model without moderator included.  $u_{00p}$  represents Level 3 deviations of pair mean intercepts from the grand mean intercept for promoted children with a value of 0 on the moderator.

Results

Table 2 presents the estimates for parameters of theoretical interest obtained from the two-piece linear growth models (see Equations 1 to 3). The first two sets of results examine the effects of retention using WJ W scores for math and reading, which implicitly compare the children’s performance to that of age-mates. The second two sets of results examine the WJ grade standard scores for math and reading, which compare the children’s performance to that of grade-mates. For ease of presentation, in the following we use  $s1_P$  and  $s2_P$  to represent Slope 1 (short term) and Slope 2 (longer term) for promoted children and  $s1_R$  and  $s2_R$  to represent Slope 1 and Slope 2 for retained children.

WJ W Scores

*WJ math W score.* As shown in Table 2, both Slope 1 and Slope 2 were positive for promoted children ( $s1_P = 16.09$ , Wald  $z = 16.67$ ,  $p < .001$ ;  $s2_P = 9.28$ , Wald  $z = 18.20$ ,  $p < .001$ ), indicating that there was an increase on WJ math W score in both pieces for promoted children, but that the increase for promoted children in the longer term was slower than the increase in the short term. Subsequent grade retention was not associated with the initial status in first grade (Wald  $z = 0.18$ , *ns*), indicating that matching on propensity scores achieved initial equivalence of the promoted and retained groups on WJ math W score. Grade retention had a negative effect on Slope 1 ( $s1_R - s1_P = -6.20$ , Wald  $z = -4.83$ ,  $p < .001$ ), indicating that in the short term, the retained children had a lower annual rate of gain of 6.20 points on the WJ math W score than did the promoted children. To put this gain in context, the normative annual gain on the WJ math W score is 9.17 for children ages 8–9. In contrast, grade retention had no significant effect on Slope 2 (Wald  $z = 1.04$ , *ns*), indicating that the annual rate of gain on the WJ math W score in the longer term did not differ significantly between the retained and promoted groups.

Coding Scheme 2 focuses on slope differences between the two pieces. Grade retention had a positive effect on the slope difference between the two pieces ( $[s2_R - s1_R] - [s2_P - s1_P] = 6.97$ , Wald  $z = 3.67$ ,  $p < .001$ ). Slope 1 and Slope 2 differed by  $-6.81$  for promoted

Table 2  
Selected Parameter Estimates for the Two-Piece Linear Growth Curve Model Without Moderators

Parameter	WJ-III math W score	WJ-III reading W score	WJ-III math grade score	WJ-III reading grade score
Intercept for promoted	459.37* (1.24)	424.70* (2.17)	98.53* (1.40)	92.33* (1.56)
Slope 1 for promoted	16.09* (0.97)	35.75* (1.66)	1.65 (1.12)	4.06* (1.21)
Slope 2 for promoted	9.28* (0.51)	12.64* (0.87)	0.01 (0.59)	-1.61* (0.64)
(Slope 2 – Slope 1) for promoted	-6.81* (1.31)	-23.11* (2.25)	-1.64 (1.52)	-5.67* (1.64)
Effect of retention on intercept	0.31 (1.73)	-1.18 (2.94)	-1.68 (1.97)	-4.01 (2.13)
Effect of retention on Slope 1	-6.20* (1.41)	-10.16* (2.40)	12.23* (1.63)	15.54* (1.75)
Effect of retention on Slope 2	0.77 (0.74)	4.77* (1.26)	-3.04* (0.86)	-2.47* (0.93)
Effect of retention on (Slope 2 – Slope 1)	6.97* (1.90)	14.93* (3.25)	-15.27* (2.20)	-18.02* (2.37)

Note. Standard errors are in parentheses. WJ-III = Woodcock-Johnson III.  
\*  $p < .05$ .

children ( $s2_P - s1_P = -6.81$ , Wald  $z = -5.20$ ,  $p < .001$ ), whereas the rate of growth did not change across the short and longer term for retained children ( $s2_R - s1_R = 0.16$ , *ns*). These results are illustrated in Figure 1A, which portrays the estimated two-piece linear growth lines for the overall retained and promoted groups.

**WJ reading W score.** Similar to the WJ math W score, there was an increase on WJ reading W score in both the short and the longer term for promoted children ( $s1_P = 35.75$ , Wald  $z = 21.54$ ,  $p < .001$ ;  $s2_P = 12.64$ , Wald  $z = 14.53$ ,  $p < .001$ ). Subsequent grade retention was not associated with initial status in first grade (Wald  $z = -0.69$ , *ns*) of the WJ reading W score. This provides further evidence of the success of the propensity score-matching procedure. Grade retention had a negative effect on Slope 1 ( $s1_R - s1_P = -10.16$ , Wald  $z = 4.23$ ,  $p < .001$ ), whereas it had a positive effect on Slope 2 ( $s2_R - s2_P = 4.77$ , Wald  $z = 3.79$ ,  $p < .001$ ), indicating that in the short term, the average annual rate gain on WJ reading W score for retained children was 10.16 points lower than that for promoted children, whereas in the longer term, retained children had an annual rate gain of 4.77 points higher on WJ reading W score than did promoted children (see Figure 1B). To put this gain in context, the normative annual gain on the WJ reading W score is 14.32 for children ages 8–9. Combining the two results using Coding Scheme 2, grade retention was found to have a positive effect on the slope difference between the two pieces ( $[s2_R - s1_R] - [s2_P - s1_P] = 14.93$ , Wald  $z = 4.59$ ,  $p < .001$ ). Slope 1 and Slope 2 differ by  $-23.11$  ( $s2_P - s1_P = -23.11$ , Wald  $z = -10.27$ ,  $p < .001$ ) for promoted children, whereas they only differed by  $-8.18$  ( $s2_R - s1_R = -8.18$ ) for retained

children. These results indicate that compared with promoted children, retained children showed a smaller reduction in slope from the short term to the longer term for WJ reading W score.

### WJ Grade Standard Scores

**WJ math grade score.** For promoted children, neither Slope 1 nor Slope 2 was significantly different from 0 ( $s1_P = 1.65$ , Wald  $z = 1.47$ , *ns*;  $s2_P = 0.01$ , Wald  $z = .02$ , *ns*). Promoted children showed nearly no change in WJ math grade standard score across 4 years. Once again, subsequent grade retention was not associated with initial status of WJ math grade score (Wald  $z = 0.85$ , *ns*). Grade retention showed a positive effect on Slope 1 ( $s1_R - s1_P = 12.23$ , Wald  $z = 7.50$ ,  $p < .001$ ), but a negative effect on Slope 2 ( $s2_R - s2_P = -3.04$ , Wald  $z = -3.53$ ,  $p < .001$ ). This finding indicates that rather than keeping a constant rate of growth relative to grademates on WJ math, as did promoted children, retained children showed a dramatic increase in the short term (the repeat year) and decreased markedly on WJ math relative to their grademates in the longer term as they encountered new material (see Figure 1C). Combining these results using Coding Scheme 2, grade retention had a negative effect on the slope difference between two pieces ( $[s2_R - s1_R] - [s2_P - s1_P] = -15.27$ , Wald  $z = -6.94$ ,  $p < .001$ ). Slope 1 and Slope 2 did not differ for promoted children ( $s2_P - s1_P = -1.64$ , Wald  $z = -1.08$ , *ns*). However, Slope 2 was significantly lower than Slope 1 for retained children ( $s2_R - s1_R = -16.91$ ).

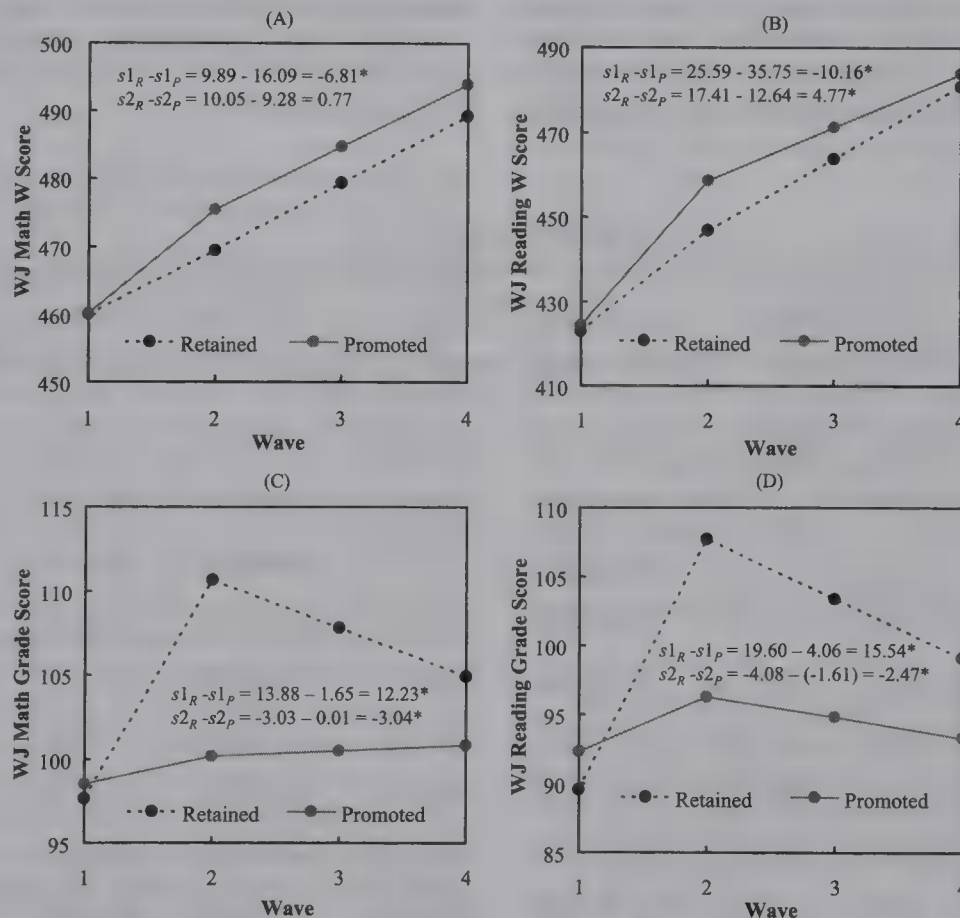


Figure 1. Estimated two-piece linear growth curves of the four WJ scores.  $s1_P$  and  $s2_P$  represent the slopes in the short and longer terms for promoted children, and  $s1_R$  and  $s2_R$  represent the slopes in the short and longer terms for retained children. WJ = Woodcock-Johnson III. \*  $p < .05$ .



*WJ reading grade score.* For promoted children, Slope 1 was positive ( $s1_P = 4.06$ , Wald  $z = 3.36$ ,  $p < .001$ ), whereas Slope 2 was negative ( $s2_P = -1.61$ , Wald  $z = -2.52$ ,  $p = .01$ ). This result indicates that for promoted children, there was an increase on WJ reading grade score in the short term followed by a slight yearly decrease on WJ reading grade score in the longer term. Again, subsequent grade retention was not associated with a significant effect on the initial status for WJ reading grade score in first grade (Wald  $z = 1.88$ , *ns*). Grade retention showed similar effects on Slope 1 and Slope 2 for WJ reading grade score as for the WJ math grade score. The effect of grade retention on Slope 1 was positive ( $s1_R - s1_P = 15.54$ , Wald  $z = 8.88$ ,  $p < .001$ ), which indicates that retained children showed an average higher annual rate gain of 15.54 points on WJ reading grade score in the short term (see Figure 1D). In contrast, the effect of grade retention on Slope 2 was negative ( $s2_R - s2_P = -2.47$ , Wald  $z = -2.66$ ,  $p = .01$ ).

Focusing now on slope differences between the two pieces using Coding Scheme 2, we found that grade retention had a negative effect on the slope difference between the two pieces ( $[s2_R - s1_R] - [s2_P - s1_P] = -18.02$ , Wald  $z = -7.60$ ,  $p < .001$ ). For promoted children, Slope 2 was 5.67 points smaller than Slope 1 ( $s2_P - s1_P = -5.67$ , Wald  $z = -3.46$ ,  $p < .001$ ). In contrast, for retained children Slope 2 was 23.69 points smaller than Slope 1 for retained children ( $s2_R - s1_R = -23.69$ ). These results show that for WJ reading grade score, retained children showed more reduction in slope from the short term to the longer term than did promoted children (see Figure 1D).

In the matched sample, there were children in the promoted group ( $n = 19$ ) who were subsequently retained in second or third grade; none of the children in the retained group were retained a second time. These results highlight the relatively high likelihood of subsequent retention for socially promoted children before the end of Grade 3 when state-mandated testing occurs. To minimize

the possible influence of retention at a later time on the effect of retention in first grade on the growth of WJ scores, we ran the same growth curve analyses with 19 pairs deleted that included the children retained after second grade. The obtained results were very similar to those reported above, and the inferences were identical on the basis of both sets of results.

### Moderating Effects

The moderating effects of the potential moderators on Slope 1 and Slope 2 for WJ W and grade scores are summarized in Table 3. For ease of presentation, we used only Coding Scheme 1. A significant moderating effect indicates that the effect of retention on Slope 1 and Slope 2 varied across different levels on a moderator. Only two of the nine moderator variables interacted with retention to affect Slope 1 for WJ reading W score and WJ math and reading grade scores. None of the moderators showed a significant moderating effect for Slope 2 for any of the WJ scores. The interaction between peer-rated externalizing problems and grade retention had a positive effect on Slope 1 for WJ reading W score ( $\gamma_{130} = 5.60$ , Wald  $z = 2.08$ ,  $p = .02$ ), indicating that in the short term, the benefit of promotion relative to retention on the WJ reading W score was higher for the children with lower levels of externalizing problems than for those with higher levels of externalizing problems (see Figure 2). The interaction between teacher-rated home-school relationship and grade retention had a negative effect on Slope 1 for WJ grade standard scores for math ( $\gamma_{130} = -8.19$ , Wald  $z = 2.05$ ,  $p = .02$ ) and for reading ( $\gamma_{130} = -8.37$ , Wald  $z = 2.04$ ,  $p = .02$ ), indicating that in the first piece, the relative benefit of retention versus promotion on grade scores was greater for children with less positive home-school relationships. Figure 3 displays the moderating effect of home-school relationship for WJ math grade score, which is very similar to that for WJ reading grade score.

Table 3

*Moderating Effects on Slope 1 and Slope 2 for Each of the Four WJ-III Scores*

Moderator	Math W score		Reading W score		Math grade score		Reading grade score	
	Slope 1 ( $\gamma_{130}$ )	Slope 2 ( $\gamma_{230}$ )	Slope 1 ( $\gamma_{130}$ )	Slope 2 ( $\gamma_{230}$ )	Slope 1 ( $\gamma_{130}$ )	Slope 2 ( $\gamma_{230}$ )	Slope 1 ( $\gamma_{130}$ )	Slope 2 ( $\gamma_{230}$ )
Age at eligibility determination	-3.97 (4.46)	2.93 (2.26)	11.96 (7.53)	-2.94 (3.80)	-8.87 (5.18)	3.67 (2.62)	5.86 (5.52)	-0.32 (2.79)
Parent-rated conduct problems	0.63 (3.27)	1.89 (1.75)	10.46 (5.60)	5.18 (3.02)	-0.24 (3.73)	-0.42 (2.00)	6.99 (3.91)	1.25 (2.11)
Parent-rated hyperactivity	0.22 (2.90)	0.47 (1.52)	-2.25 (5.01)	0.53 (2.65)	-0.78 (3.32)	-0.37 (1.74)	-3.30 (3.51)	-0.83 (1.84)
Teacher-rated conduct problems	2.27 (2.79)	1.19 (1.50)	3.98 (4.54)	2.54 (2.42)	0.08 (3.40)	1.39 (1.82)	0.85 (3.42)	2.19 (1.83)
Teacher-rated hyperactivity	1.42 (2.47)	0.75 (1.31)	2.49 (4.04)	3.05 (2.14)	2.11 (3.00)	-0.59 (1.60)	2.38 (3.04)	1.17 (1.61)
Peer-rated externalizing problems	2.26 (1.67)	0.82 (.88)	5.60* (2.69)	-0.97 (1.43)	-0.38 (1.91)	0.90 (1.01)	1.82 (2.00)	0.09 (1.06)
Teacher-student relationship	0.96 (1.96)	-0.90 (1.19)	-3.68 (3.28)	0.36 (1.97)	4.10 (2.23)	-1.72 (1.37)	0.43 (2.46)	-.28 (1.46)
Home-school relationship	-3.49 (3.28)	1.70 (1.73)	-8.35 (5.37)	3.00 (2.81)	-8.19* (4.00)	4.05 (2.12)	-8.37* (4.10)	3.78 (2.15)
Resilient personality	-0.15 (1.75)	-1.05 (.95)	-5.17 (2.88)	-0.49 (1.55)	1.63 (2.00)	-0.18 (1.08)	-2.48 (2.15)	0.36 (1.16)

*Note.* Standard errors are in parentheses. Moderating effects are indicated by the coefficients associated with the interaction between a certain moderator and retention status following grade 1 ( $\gamma_{130}$  for Slope 1 and  $\gamma_{230}$  for Slope 2; see Equations 4 and 5). WJ-III = Woodcock-Johnson III.

\*  $p < .05$ .

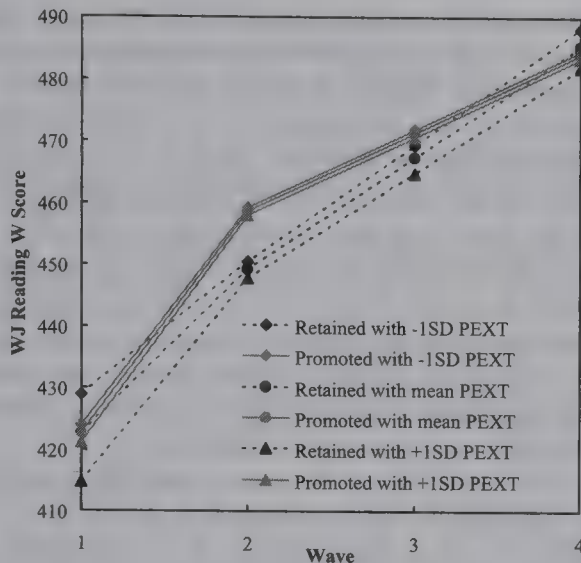


Figure 2. Estimated two-piece linear growth curves of WJ reading W score for retained and promoted children with different levels of peer-rated externalizing problems (PEXT). WJ = Woodcock-Johnson III.

## Discussion

In this study, we used two-piece linear growth curve models to examine the short-term and longer term effects of grade retention on the change of WJ math and reading W and grade scores across 4 years. Analyses were conducted with retained and promoted children who were closely matched using modern propensity-matching procedures. One indicator of success of the propensity-matching procedure is the lack of initial differences in achievement between the retained and promoted children (see also Table 1).

### W Scores

For W scores, grade retention decreased the growth rate in the first piece (short term), but had either no significant effect on the growth rate (math) or increased the growth rate (reading) in the second piece (longer term). Thus, retained children grew more slowly in reading in the repeat year than did matched promoted children but “caught up” to the promoted children in the longer term. In the longer term, they grew in reading at a faster rate than did their matched promoted peers such that by Time 4 (when retained children were in third grade and most promoted children were in fourth grade), retained and promoted children did not differ significantly in reading ability.<sup>3</sup> Although it is tempting to extrapolate the linear growth for Slope 2 beyond Time 4, such extrapolation would be risky for related statistical and substantive reasons. Statistically, given the limited number of measurement waves collected so far in the present study, we were not able to fit nonlinear models of longer term growth. It is likely that longer term change will approach an asymptote over time so that linear growth only provides an approximation of the true form of growth over the 4-year period that was studied. Extrapolation beyond the range of the data from models that are only approximate is risky (Cook, 1993; MacCallum, 2003). Substantively, there is a rather abrupt change in the nature of reading and reading instruction that occurs between Grades 3 and 4 (Chall, 1996; Duke & Pearson, 2002). As retained children make this transition, their growth may slow. For math, the faster growth for promoted children in the short term

disappears in the longer term when retained and promoted children’s rate of growth is equivalent. However, the early disadvantage of retention on math growth is not compensated by more rapid growth in the longer term.

### Grade Standard Scores

In contrast to results for W scores, grade retention increased the growth rates of both the WJ math and reading grade standard scores in the short term but decreased growth rate of both scores in the longer term. Whereas promoted children showed only a slight decrease in their rate of growth in math and a somewhat larger decrease in their rate of growth in reading, retained children showed large drops in their growth rate from the first piece to the second piece for both the math and the reading grade standard scores. Thus, children did benefit from grade retention in the short term in terms of their performance relative to national grade norms. Previously published research on this sample at the end of Measurement Wave 2 found that retained students were rated by their classroom teacher as achieving more in the classroom than did their low-achieving promoted peers (Gleason, Kwok, & Hughes, 2007). The current findings suggest that this benefit erodes as retained students encounter an unfamiliar and more challenging curriculum. Again, additional waves of data are needed to determine whether the short-term benefits of retention on grade standard scores evaporate or reverse with additional years postretention.<sup>4</sup> Of particular concern is the impact of the sequence of failure, success, and failure that retained children appear to experience in their first 4 years of formal schooling on their academic self-efficacy and beliefs regarding the degree to which they can control academic outcomes. In future studies, we will investigate whether the documented association between grade retention and school withdrawal is mediated, in part, by the effect of grade retention on such academic-related beliefs.

### Moderators

We investigated nine potential moderators in a total of 72 tests (9 moderators  $\times$  4 achievement outcomes  $\times$  2 pieces). By chance alone, 3.6 tests would be expected to be statistically significant at an alpha of less than .05. Thus, the three significant results obtained should be interpreted as preliminary and requiring replication. As expected, the relative short-term benefit of promotion versus retention on growth in reading ability was greater for children with lower levels of peer-rated externalizing problems. This finding is consistent with the view that children with few behavioral problems (and greater prosocial skills) may be better

<sup>3</sup> Difference =  $-1.50$ , Wald  $z = -0.51$ ,  $p = .31$ .

<sup>4</sup> This study makes comparisons across grades on the basis of extensive normative data underlying the Woodcock-Johnson III grade standard scores. At this point in the longitudinal study, direct comparison of the trajectories of the retained and promoted groups (same-grade comparison) is not possible in the two-piece growth model that we used. Such a comparison would require discarding the measurement taken in the repeated year for the retained participants. Such a procedure would leave us with only two waves of postretention measurement (second grade and third grade) on which to compare the two groups. At least three postretention measurements are required to estimate the two-piece model of growth.



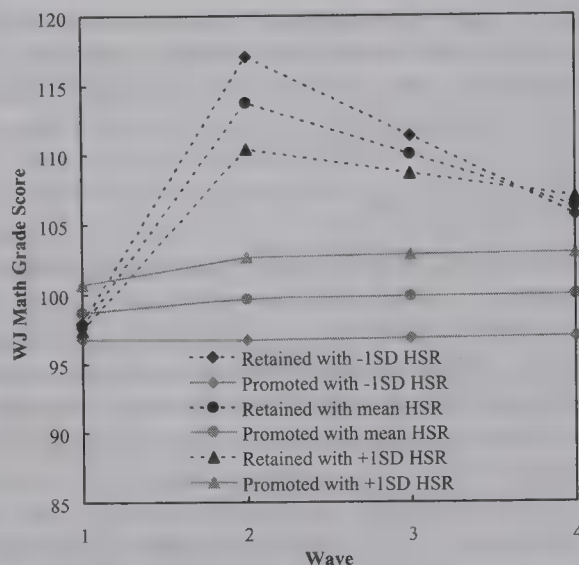


Figure 3. Estimated two-piece linear growth curves of WJ math grade score for retained and promoted children with different levels of home-school relationship (HSR). WJ = Woodcock-Johnson III.

prepared to meet the challenges of a more challenging curriculum. Similarly, the relative short-term benefit of retention versus promotion on growth in grade standard scores is greater for children who do not have strong parent involvement in their education. The limited number of significant interactions between retention and the moderator variables obtained in this study does suggest that the effects of retention are robust across a range of individual differences. An alternative explanation, of course, is that we failed to select the most appropriate variables for our tests of moderation.

### Strengths and Limitations of the Present Study

In contrast to many of the previous investigations of grade retention, this investigation drew on several recent advances in the design and analysis of observational studies that can greatly strengthen causal inferences (see West & Thoenes, 2008, for a review). First, we selected a large sample of children at risk for retention before the retention decision. This feature was expected to reduce, but not eliminate, baseline differences between retained and promoted children. It maximized the overlap between the promoted and the retained groups, eliminating the need to use statistical adjustment methods that may involve extrapolation. Second, we used an extensive battery of 72 pretest measures that were expected to be related to retention, math and reading achievement scores, or both. These measures were used to create propensity scores that were used to optimally match the retained and nonretained children before retention. Rosenbaum and Rubin (1983; see also Rosenbaum, 2002) showed analytically that propensity score methods provide a device that properly equates groups in observational studies if all of the important variables have been measured and balance on the variables is achieved between the treatment and control groups. Our selection of 72 baseline variables based on current theory and research represented a thorough attempt to measure all important baseline variables. As shown in Table 1, remaining differences between the groups after matching on propensity scores were minimal and nonsignificant. Third, we conducted growth curve modeling of the

WJ achievement scores that are scaled by means of the Rasch model. Rasch model scaling yields interval-level measurements on a single dimension. The use of the WJ measures that are not part of school district or state education assessments prevents teachers from "teaching to the test." Growth modeling permitted comparison of the slopes of the retained and promoted groups, both in the short term and in the longer term. This method represents an alternative strategy of comparing the groups. The combination of the use of two distinct methods, propensity scores and growth modeling, gave us two chances to get the adjustment for preexisting group differences right, giving us far greater confidence although not certainty in the causal interpretation of the effects of grade retention in the present study.

The primary limitations of the present study result are associated with the collection of only four waves of data to date. First, the number of waves of data restricts the forms of growth that could be investigated. We were able to investigate both short- and longer term growth in both WJ W scores and WJ grade standard scores. However, with only four waves of observation, our modeling was limited to the examination of linear effects. Other possible nonlinear models, such as the effects of grade retention increasing or decreasing over time to an asymptote, could not be investigated. We are currently collecting additional waves of data that will permit examination of such nonlinear effects. Second, additional waves of data collection will also permit a direct comparison between matched retained and promoted children when they are in the same grade. Same-grade comparisons involve examining the mean achievement of retained and promoted students when they are in the same grade, but not in the same year. Thus, the measurement of achievement for the retained children typically lags 1 year behind that for the promoted peers. Such direct grade comparisons are important because educators and parents may be more concerned with how retained students perform relative to their current classmates and grade expectations than to their former classmates, who are in a different grade (Lorenz, 2006).

In summary, the picture is complicated. Results differ on the basis of the scale used (age or grade), time elapsed since retention year, and achievement domain (reading vs. math). These results suggest that the question "What is the effect of grade retention on achievement?" is inappropriate. The more appropriate question is "What is the effect of grade retention, for whom, on what academic competencies, in reference to what standard (age or grade), at what point in time postretention?" With additional waves of data, we hope to provide a more complete picture of the effects of retention on achievement over time.

### References

- Alexander, K. L., Entwistle, D. R., & Dauber, S. L. (1994). *On the success of failure: A reassessment of the effects of retention in the primary grades*. Cambridge, England: Cambridge University Press.
- Alexander, K. L., Entwistle, D. R., & Dauber, S. L. (2003). *On the success of failure: A reassessment of the effects of retention in the primary grades* (2nd ed.). Cambridge, England: Cambridge University Press.
- Bracken, B. A., & McCallum, R. S. (1998). *Universal Nonverbal Intelligence Test: Examiner's manual*. Itasca, IL: Riverside.
- Buhrmester, D., & Furman, W. (1987). The development of companionship and intimacy. *Child Development*, 54, 1386-1399.
- Caspi, A., Block, J., Block, J. H., & Klopp, B. (1992). A "common-

- language" version of the California Child Q-set for personality assessment. *Psychological Assessment*, 4, 512–523.
- Chall, J. S. (1996). *Stages of reading development* (2nd ed.). Fort Worth, TX: Harcourt Brace.
- Cicchetti, D., & Posner, M. I. (2005). Cognitive and affective neuroscience and developmental psychopathology. *Development and Psychopathology*, 17, 569–575.
- Cillessen, A. H. N., & Bukowski, W. M. (2000). *Recent advances in the measurement of acceptance and rejection in the peer system*. San Francisco: Jossey-Bass.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Cook, T. D. (1993). A quasi-sampling theory of the generalization of causal relationships. In L. B. Sechrest & A. G. Scott (Eds.), *New directions for program evaluation* (No. 57, 39–81). San Francisco: Jossey-Bass.
- Dennebaum, J. M., & Kulberg, J. M. (1994). Kindergarten retention and transition classrooms: Their relationship to achievement. *Psychology in the Schools*, 31, 5–12.
- Dickey, W. C., & Blumberg, S. J. (2004). Revisiting the factor structure of the Strengths and Difficulties Questionnaire: United States, 2001. *Journal of the American Academy of Child & Adolescent Psychiatry*, 43, 1159–1167.
- Duke, N. K., & Pearson, P. D. (2002). Effective practices for developing reading comprehension. In A. E. Farstrup & S. J. Samuels (Eds.), *What research has to say about reading instruction* (3rd ed., pp. 205–242). Newark, DE: International Reading Association.
- Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling*, 8, 430–457.
- Evans, W., Baugh, C., & Sheffer, J. (2005). A study of the sustained effects of comprehensive school reform programs in Pennsylvania. *School Community Journal*, 15, 15–28.
- Gleason, K. A., Kwok, O., & Hughes, J. N. (2007). The short-term effect of grade retention on peer relations and academic performance of at-risk first graders. *Elementary School Journal*, 107, 327–340.
- Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, 38, 581–586.
- Goodman, R. (2001). Psychometric properties of the Strengths and Difficulties Questionnaire. *Journal of Child Psychology and Psychiatry*, 40, 791–799.
- Gootman, E. (2005, March 19). One in 3 city 4th graders may not advance to 5th. *New York Times*, Section B, p. 5.
- Hill, C. R., & Hughes, J. N. (2007). An examination of the convergent and discriminant validity of the Strengths and Difficulties Questionnaire. *School Psychology Quarterly*, 22, 380–406.
- Holmes, C. T. (1989). Grade-level retention effects: A meta-analysis of research studies. In L. A. Shepard & M. L. Smith (Eds.), *Flunking grades: Research and policies on retention* (pp. 16–33). London: Falmer Press.
- Holmes, C. T., & Matthews, K. M. (1984). The effects of nonpromotion on elementary and junior high school pupils: A meta-analysis. *Review of Educational Research*, 54, 225–236.
- Hong, G., & Raudenbush, S. W. (2005). Effects of kindergarten retention policy on children's cognitive growth in reading and mathematics. *Educational Evaluation and Policy Analysis*, 27, 205–224.
- Hughes, J. (1990). Assessment of children's social competence. In C. R. Reynolds & R. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children* (pp. 423–444). New York: Guilford Press.
- Hughes, J. N., Cavell, T. A., & Willson, V. (2001). Further evidence of the developmental significance of the teacher-student relationship. *Journal of School Psychology*, 39, 289–302.
- Hughes, J. N., & Kwok, O. (2006). Classroom engagement mediates the effect of teacher-student support on elementary students' peer acceptance: A prospective analysis. *Journal of School Psychology*, 43, 465–480.
- Huitema, B. E. (1980). *The analysis of covariance and alternatives*. New York: Wiley.
- Jimerson, S. R. (1999). On the failure of failure: Examining the association between early grade retention and education and employment outcomes during late adolescence. *Journal of School Psychology*, 37, 243–272.
- Jimerson, S. R. (2001). Meta-analysis of grade retention research: Implications for practice in the 21st century. *School Psychology Review*, 30, 420–437.
- John, O. P., & Srivastava, S. (1999). The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 102–138). New York: Guilford Press.
- Kwok, O., Hughes, J. N., & Luo, W. (2007). The role of personality resilience on lower achieving first grade students' current and future achievement. *Journal of School Psychology*, 45, 61–82.
- Lerner, R. M. (1989). Developmental contextualism and the life-span view of person-context interaction. In M. H. Bornstein & J. S. Bruner (Eds.), *Interaction in human development. Crosscurrents in contemporary psychology* (pp. 217–239). Hillsdale, NJ: Erlbaum.
- Little, R. J., Hyonggin, J., Johanns, J., & Giordani, B. (2000). A comparison of subset selection and analysis of covariance for the adjustment of confounders. *Psychological Methods*, 5, 459–476.
- Lorence, J. (2006). Retention and academic achievement research revisited from a United States perspective. *International Education Journal*, 7, 731–777.
- MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research*, 38, 113.
- Mantzicopoulos, P. (2003). Flunking kindergarten after head start: An inquiry into the contribution of contextual and individual variables. *Journal of Educational Psychology*, 95, 268–278.
- Mantzicopoulos, P., & Morrison, D. (1992). Kindergarten retention: Academic and behavioral outcomes through the end of the second grade. *American Educational Research Journal*, 29, 182–198.
- McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9, 403–425.
- McCoy, A. R., & Reynolds, A. J. (1999). Grade retention and school performance: An extended investigation. *Journal of School Psychology*, 37, 273–298.
- Meehan, B. T., Hughes, J. N., & Cavell, T. A. (2003). Teacher-student relationships as compensatory resources for aggressive children. *Child Development*, 74, 1145–1157.
- Miesels, S. J., & Liaw, F. R. (1993). Failure in grade: Do retained students catch up? *Journal of Educational Research*, 8, 69–77.
- Miles, S. B., & Stipek, D. (2006). Contemporaneous and longitudinal associations between social behavior and literacy achievement in a sample of low-income elementary school children. *Child Development*, 77, 103–117.
- Ming, K., & Rosenbaum, P. A. (2001). A note on optimal matching with variable controls using the assignment algorithm. *Journal of Computational and Graphical Statistics*, 10, 455–463.
- No Child Left Behind Act of 2001, P. L. 107–110. (2001).
- Pagani, L., Tremblay, R. E., Vitaro, F., Boulerice, B., & McDuff, P. (2001). Effect of grade retention on academic performance and behavioral development. *Development and Psychopathology*, 13, 297–315.
- Pianta, R. C., Tietohl, P. J., & Bennett, E. M. (1997). Differences in social adjustment and classroom behavior between children retained in kinder-



- garten and groups of age and grade matched peers. *Early Education and Development*, 8, 137–152.
- Pierson, L. H., & Connell, J. P. (1992). Effects of grade retention on self-system processes, school engagement, and academic performance. *Journal of Educational Psychology*, 84, 300–307.
- Realmuto, G. M., August, G. J., Sieler, J. D., & Pessoa-Brandao, L. (1997). Peer assessment of social reputation in community samples of disruptive and nondisruptive children: Utility of the Revised Class Play Method. *Journal of Clinical Child Psychology*, 26, 67–76.
- Reichardt, C. S. (2006). The principle of parallelism in the design of studies to estimate treatment effects. *Psychological Methods*, 11, 1–18.
- Reynolds, A. J. (1992). Grade retention and school adjustment: An explanatory analysis. *Education Evaluation and Policy Analysis*, 14, 101–121.
- Reynolds, A. J., & Bezruczko, N. (1993). School adjustment of children at risk through fourth grade. *Merrill-Palmer Quarterly*, 39, 457–480.
- Roderick, M. (1994). Grade retention and school dropout: Investigating the association. *American Educational Research Journal*, 31, 729–759.
- Roderick, M., Bryk, A. S., Jacob, B. A., Easton, J. Q., & Allensworth, E. (1999). *Ending social promotion: Results from the first two years*. Chicago: Consortium on Chicago School Research. Retrieved August 9, 2002, from [www.consortium-chicago.org/publications/p0g04.html](http://www.consortium-chicago.org/publications/p0g04.html)
- Roderick, M., & Nagaoka, J. (2005). Retention under Chicago's high stakes testing program: Helpful, harmful, or harmless? *Educational Evaluation and Policy Analysis*, 27, 309–340.
- Rosenbaum, P. A. (2002). *Observational studies* (2nd ed.). New York: Springer.
- Rosenbaum, P. A., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41–55.
- Sameroff, A. J. (1975). Transactional models in early social relations. *Human Development*, 18, 65–79.
- Sameroff, A. J. (1989). Principles of development and psychopathology. In A. Sameroff & R. Emde (Eds.), *Relationship disturbances in early childhood* (pp. 17–32). New York: Basic Books.
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177.
- Shadish, W. R., & Clark, M. H. (2006, July). *A randomized experiment comparing random to nonrandom assignment*. Paper presented at the Symposium on Causality 2006, Jena, Germany. Retrieved March 11, 2007, from <http://www.metheval.uni-jena.de/projekte/symposium2006/contributions.php>
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.
- Shadish, W. R., Luellen, J. K., & Clark, M. H. (2006). Propensity scores and quasi-experiments: A testimony to the practical side of Lee Sechrest. In R. R. Bootzin & P. E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp. 143–157). Washington, DC: American Psychological Association.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.
- Sipple, J. W., Killeen, K., & Monk, D. H. (2004). Adoption and adaptation: School district responses to state imposed learning and graduation requirements. *Educational Evaluation and Policy Analysis*, 26, 143–168.
- Skinner, E. A., Zimmer-Gembeck, M. J., & Connell, J. P. (1998). Individual differences and the development of perceived control. *Monographs of the Society for Research in Child Development*, 63(Serial No. 254, Nos. 2–3).
- Terry, R. (1999, April). *Measurement and scaling issues in sociometry: A latent trait approach*. Paper presented at the biennial meeting of the Society for Research in Child Development, Albuquerque, New Mexico.
- Texas Education Agency. (2005). *Grade-level retention in Texas public schools, 2003–04* (Document No. GE06 601 01). Austin, TX: Author.
- Warren, J. R., & Edwards, M. R. (2005). High school exit examinations and high school completion: Evidence from the early 1990s. *Educational Evaluation and Policy Analysis*, 27, 53–74.
- West, S. G., & Thoemmes, F. (2008). Equating groups. In P. Alasuutari, L. Bickman, & J. Brannon, (Eds.), *Handbook of social research methods* (pp. 414–430). London: Sage.
- Willson, V., & Hughes, J. N. (2006). Retention of Hispanic/Latino students in first grade: Child, parent, teacher, school, and peer predictors. *Journal of School Psychology*, 44, 31–49.
- Wong, S. W., & Hughes, J. N. (2006). Ethnicity and language contributions to dimensions of parent involvement. *School Psychology Review*, 35, 645–662.
- Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson tests of achievement: Standard and supplemental batteries*. Allen, TX: DLM Teaching Resources.
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *WJ-III Tests of Achievement*. Itasca, IL: Riverside.
- Woodcock, R. W., & Muñoz-Sandoval, A. F. (1993). *Woodcock-Muñoz Language Survey*. Chicago: Riverside.
- Woodcock, R. W., & Muñoz-Sandoval, A. F. (1996). *Batería Woodcock-Muñoz Pruebas de Aprovechamiento—Revisada*. Itasca, IL: Riverside.
- Woodcock, R. W., & Muñoz-Sandoval, A. F. (2001). *Comprehensive manual: Woodcock-Muñoz Language Survey Normative Update*. Itasca, IL: Riverside.
- Wu, W., West, S. G., & Hughes, J. N. (2008). Short-term effects of grade retention on growth rate of Woodcock-Johnson III broad math and reading scores. *Journal of School Psychology*, 46, 85–105.

Received April 19, 2007

Revision received December 6, 2007

Accepted December 17, 2007 ■

# Classroom Age Composition and Developmental Change in 70 Urban Preschool Classrooms

Arlen C. Moller

Children's Institute, Inc. and Gettysburg College

Emma Forbes-Jones and A. Dirk Hightower

Children's Institute, Inc. and University of Rochester

A multilevel modeling approach was used to investigate the influence of age composition in 70 urban preschool classrooms. A series of hierarchical linear models demonstrated that greater variance in classroom age composition was negatively related to development on the Child Observation Record (COR) Cognitive, Motor, and Social subscales. This was true when controlling for class size, general classroom quality, and socioeconomic status at the classroom level and for age, gender, and baseline ability at the child level. Additionally, to address possible concerns related to nonrandom assignment to classrooms, a series of models were run including variance in developmental age (i.e., baseline ability) at the classroom level and at the child level. The results were consistent for chronological age composition and developmental age composition at the classroom level; greater variance in classroom developmental age composition was negatively related to Time 2 scores on the COR Cognitive, Motor, and Social subscales. Furthermore, a cross-level interaction indicated that negative influence of greater variance in classroom developmental age composition was stronger for children older in developmental age. Implications for early childhood education policy are discussed.

**Keywords:** preschool, mixed age, single age, age composition, Child Observation Record

There is a great deal of variability in the age composition of U.S. preschool classrooms. Some preschool programs tightly restrict the range of ages in their classrooms (e.g., 3-year-olds only). Other programs permit, and in some cases even promote, a wide range of ages in the same classroom (e.g., from 2.5 years to as old as 5 years). Such variability in practice follows from a lack of consensus in both the theoretical and the empirical literature on the effects of age composition of classrooms on individual learning.

Research on the effects of classroom age composition on individual growth is not new. The theoretical writings of two pioneers in the field of early childhood development, Vygotsky and Piaget, are at odds with respect to the predicted impact of exposure of individual children to peers of variant ages on cognitive and social development (Bailey, Burchinal, & McWilliams, 1993; Hartup, 1993; Piaget, 1932; Rubin, Bukowski, & Parker 1998; Vygotsky,

1930/1978). From one perspective, that of Piaget (1932) and others, interacting with peers who are close in age is preferable. Children close in age are also likely to be more closely matched in terms of knowledge, skill, and power. As a result, children in such a setting are able to engage in and resolve cognitive and social conflicts with each other and play a variety of roles in their resolution. By contrast, Vygotsky and others argued that interaction with older, more competent children provides an optimal context for cognitive and behavioral development (Vygotsky, 1930/1978).

Interestingly, in the years since, relatively little empirical consensus has emerged. Yet despite a lack of consensus, the National Association for the Education of Young Children has given mixed-age classrooms its endorsement<sup>1</sup> (Bredekamp, 1987), in line with several researchers in the field (e.g., Katz, Evangelou, & Hartman, 1990; Whaley & Kantor, 1992). We now explore the extant literature relating to preschool classroom age composition more closely.

## Support for Mixed-Age Preschool Classrooms

Proponents of mixed-age preschool classrooms typically focus on the age-specific benefits most relevant to the younger children in such groupings. For example, as already noted, Vygotsky (1930/1978) spoke of the (perhaps) optimal challenges inherent in classroom contexts that include children who are older and thus more mature. Similarly, social learning theorists have maintained that

---

Arlen C. Moller, Department of Psychology, Gettysburg College, and Children's Institute, Inc., Rochester, New York; Emma Forbes-Jones, Department of Psychiatry, University of Rochester School of Medicine and Dentistry, and Children's Institute, Inc.; A. Dirk Hightower, Department of Clinical and Social Sciences in Psychology, University of Rochester, and Children's Institute, Inc.

We would like to thank those responsible for managing the Rochester Early Childhood Assessment Partnership Project, including Lauri Brugger, Susan Greenberg, Andrew MacGowan, Beverly Miller, and Kathy Slater. In particular, we thank Walter Gramiak for his help and expertise in managing and cleaning the dataset from which these analyses were derived. Finally, we thank Guillermo Montes, Aaron Metzger, and Janis Cameron for their insightful comments and feedback on earlier versions of this article.

Correspondence concerning this article should be addressed to Arlen C. Moller, Department of Psychology, Gettysburg College, Box 0407, 300 North Washington Street, Gettysburg, PA 17325. E-mail: amoller@gettysburg.edu

---

<sup>1</sup> The National Association for the Education of Young Children's (1996) most recent position statement does not take a position with regard to the age composition of preschool classrooms, and the recommendation for mixed-age classrooms extended by Bredekamp (1987) has not been retracted.



younger children benefit from being exposed to and modeling the behavior of older children (Bandura, 1986). In support, Rothstein-Fisch and Howes (1988) found that when toddlers interacted with slightly older peers, the younger children exhibited more complex language and engaged in more complex play in a family childcare setting. Howes and Farver (1987) similarly found that toddlers whose play partners were older peers engaged in more complex social play.

Other advocates for mixed-age classrooms have posited that older children, too, may benefit from exposure to younger children (Katz et al., 1990; Lloyd, 1999). For example, in such settings older children are in a position to more readily practice prosocial behavior (Derscheid, 1997; Urberg & Kaplan, 1986) and leadership skills (French, Waas, Stright, & Baker, 1986), given their elevated status in mixed-age group interactions. Chapman (1994) found that older children in a mixed-age elementary school setting served as mentors for younger children both spontaneously and when asked. Although the specific benefits that have been linked to mixed-age peer interaction have thus far been age specific (i.e., only relevant to younger or older children), advocates of mixed-age classrooms maintain that the age-specific advantages are presumably passed on as a child progresses through such a program (i.e., younger children will one day be older children in the same classroom), and thus an egalitarian ethic is preserved.

Further research has suggested additional benefits of mixed-age classrooms for all children, regardless of age. For example, research by Goldman (1981) has suggested that both 3- and 4-year-olds in mixed-age preschool classrooms engaged in significantly less parallel (as opposed to interactive) forms of play, relative to those in same-age classrooms. An investigation by Urberg and Kaplan (1986) found that both older and younger children in mixed-age preschool classrooms exhibited greater interactive functional play and less parallel functional play, relative to those in same-age classrooms. Rubin, Maioni, and Hornung (1976) have argued that interactive play represents a more mature level of play, relative to parallel play. Thus, these authors argued that mixed-age classrooms facilitate more mature play and therefore more optimal social-cognitive development.

Whaley and Kantor (1992) suggested that mixed-age classrooms may help to lessen the likelihood of parents and caregivers to make comparisons among children in terms of development. Instead, mixed-age rooms allow parents to focus on each child's development as an individual. As such, mixed-age classrooms may decrease competition between peers, focusing children instead on more mastery-oriented goals. Many studies conducted with children, adolescents, and college students confirm that focusing on individual standards (i.e., mastery goals) is optimal in terms of maintaining both intrinsic motivation and achievement in the classroom (see Elliot & Dweck, 2005, and Nicholls, 1989, for reviews). Blasco, Bailey, and Burchinal (1993) found that children with developmental disabilities, in particular, were more likely to engage in mastery play when they were enrolled in mixed-age settings.

Whaley and Kantor (1992) further speculated that benefits extend to both children and teachers as a result of the "family climate" and continuity of care fostered by mixed-age classrooms. That is, these authors suggested that the abrupt transitions in climate encountered when moving between restricted-age classrooms at larger childcare facilities could be stressful for both the

teachers and the children. However, an empirical investigation into the perceived stress levels of children and teachers in such facilities has yet to be published.

Several authors have noted that gender segregation is less prevalent in mixed-age classrooms relative to restricted-age settings (Field, 1982; Lougee, Grueneich, & Hartup, 1977; Roopnarine et al., 1992; Roopnarine & Johnson, 1984). It seems that in mixed-age contexts, age plays a more powerful role than gender in peer playmate selection. Thus, at a society level, long-term promotion of gender equality might be better served by more preschools moving toward a greater diversity in terms of classroom age composition. Such an argument is highly speculative, as no empirical explorations into a relation between preschool classroom age composition and attitudes related to gender equality have been conducted.

Finally, some research has supported the notion that mixed-age peer interactions occur naturally, outside the classroom context (Ellis, Rogoff, & Cromer, 1981). An observational study by Ellis et al. (1981) of 436 children ages 1–2 found that target children were with same-age peers in only 6% of the observations recorded, compared with 55% of the observations involving mixed-age peer interactions (with children who differed in age by at least 1 year). These findings related to naturalistic social interactions, however, are inconclusive to the extent that this may be because of peer accessibility or children actively seeking out companionship from mixed-age peers. Goldman's (1981) research on mixed- versus same-age preschool classrooms demonstrated that—within mixed-age classrooms—age did not significantly influence choice of playmates. Thus, some authors have suggested that children are naturally better suited to developing in the context of mixed-age groups (Ellis et al., 1981; Whaley & Kantor, 1992).

### Support for Restricted-Age Preschool Classrooms

The central argument in favor of restricting the range of age in preschool classrooms concerns a desire to provide each child with an optimal degree of challenge and the utility of specialization toward meeting this goal. As such, a criticism that has been levied against mixed-age classrooms in favor of more restricted-age preschool groupings concerns the development of age-appropriate curriculum and staff training. Specifically, when teacher training and curriculum are uniquely focused on what is developmentally appropriate for a more restricted age range of children, the quality of the classroom climate will be higher.

Although some authors have maintained that mixed-aged classrooms facilitate more mature forms of play (Goldman, 1981; Howes & Farver, 1987; Rothstein-Fisch & Howes, 1998; Urberg & Kaplan, 1986), other researchers have found evidence to the contrary or suggestive of a more complex effect than is often highlighted. For example, Roopnarine et al. (1992) observed more social and nonsocial dramatic play in preschoolers attending same-age relative to mixed-age classrooms, whereas more social-constructive and nonsocial manipulative play was seen in mixed-age classrooms. Roopnarine et al. maintained that dramatic play was the more complex form of play relative to manipulative play, suggesting that same-age classrooms facilitated more sophisticated play behaviors. Urberg and Kaplan (1986) reported finding a greater incidence of interactive, mature forms of play in mixed-age preschool classrooms relative to same-age classrooms; however,



these authors also reported several observed costs associated with mixed-age classrooms. For instance, younger children in mixed-age classrooms tended to be involved in a greater number of negative interactions. This was attributed to the potential for frustration in older children resulting from interactions with less mature younger children. Furthermore, Urberg and Kaplan found that older children in mixed-age classrooms were more likely to be onlookers, detaching themselves from social interaction. Similarly, Goldman (1981) found more solitary play and less participation in teacher-directed activities among older children in mixed-age classrooms. Urberg and Kaplan speculated that teachers in the mixed-age classrooms might have had a more difficult time providing a sufficient variety of activities suitable for keeping children of varied age and ability interested and involved.

Proponents of mixed-age classrooms often focus on benefits to younger children, but the potential costs to the older children in such classrooms may sometimes be overlooked. A study of mixed-age elementary school classrooms (children ages 6 to 9 years) by Byrnes, Shuster, and Jones (1994) confirmed that both older students and their parents rated the quality of the academics in multiage classrooms less positively than that in single-age classrooms. Despite failing to report any benefits found for older, more able students in their study, the authors suggest that administrators ought to pay special attention to helping "more able students in seeing how they benefit from being in multiage classrooms" (p. 21). In another observational investigation of play behavior in different classrooms, Dunn, Kontos, and Potter (1996) monitored family childcare homes and found that frequency of interactions with older peers were associated with more complex cognitive play, whereas interactions with younger and same-age peers were associated with less complex social and cognitive play and lower receptive language scores. One lab experiment (Tudge, 1992) compared mixed- versus matched-competence peers (ages 5–9 years) working in dyads. The results of this experiment suggested that regression in thinking was as likely a consequence of mixed groupings as was improvement. These results suggest that at least with regard to some cognitive outcomes, older children in mixed-age groups may suffer. Furthermore, support for such a hypothesis was found by Winsler et al. (2002). Winsler et al. studied longitudinally one preschool's transition from two single-age classrooms (separate 3- and 4-year-olds) to two mixed-age classrooms (combined 3- and 4-year-olds). In this study, statistical interactions were found between child age and classroom age composition in predicting task focus in the classroom. Although 3-year-olds seemed to do better after the change to mixed-age classrooms, 4-year-olds were observed to be less focused and goal directed. Moreover, children of both ages were observed demonstrating significantly more overt positive affect (e.g., smiles and laughter) and less neutral affect in the single-age classrooms compared with the mixed-aged classrooms.

The 1995 meta-analysis conducted by Veenman concluded that there were no academic or social differences between elementary school children attending mixed-age and same-age classrooms. Similarly, Blasco et al. (1993) reported no differences in terms of complex mastery play in same- versus mixed-age classrooms when their observations focused on typically developing preschool-age children; the benefits attributed to mixed-age classrooms and reported above were (again) specific to children with developmental disabilities.

## Ability-Based Groupings

As noted, many of the theoretical assertions for or against mixed- and restricted-age groupings are based on the assumption that age is closely related to ability. By this token, ability-based groupings may be expected to have an influence on developmental outcomes similar to that of age-based groupings. Mirroring the literature supporting the efficacy of mixed-age classrooms, many researchers have championed mixed-ability groupings in terms of benefits to be reaped by disadvantaged children who are lower in ability (Guralnick, 1981; Rogers & Ross, 1986; Tach & Fargas, 2006). For example, Tach and Fargas (2006) found that within kindergarten and first-grade classrooms in which ability-based groupings were used, placement into a higher ability group exerted a uniformly positive effect on student learning-related behavior and reading achievement. Typically, these studies have implicitly assumed that the costs for higher ability children in such mixed-ability groupings are negligible, although several exceptions (e.g., Guralnick, 1981) have reported no detrimental impact on higher ability children in mixed-ability groupings. Nevertheless, a theme that runs through these studies advocating for mixed-ability groupings is that an important and focal goal is ensuring that the needs of students who have experienced situations that place them at risk or are disadvantageous for their academic development are met. From an entirely different perspective, some researchers have advocated expressly for the interests of the most gifted and talented children. Researchers focused on those children higher in ability tend to advocate for more restricted-ability groupings, such that high-ability children have more opportunities to interact with peers of similar ability (e.g., Gallagher, 1986). To summarize, as is the case in the literature focused on age-related groupings, the literature on ability-related groupings is mixed; furthermore, the most controversial aspect of the debate concerns those children in mixed contexts who are older, higher in ability, or both.

## Shortcomings of the Extant Literature Addressed by the Present Investigation

Many of the studies that have expressly focused on classroom age composition have targeted an elementary school population, which makes it difficult to extend the conclusions to preschool classrooms (e.g., Chapman, 1994; French et al., 1986; Veenman, 1995). Early childhood (ages 2 to 5) is a period of relatively rapid development; thus, the impact of age differences in classroom composition within this range is likely to be magnified, if not distinctly different, in preschool classrooms relative to that found in elementary school settings among older children.

Furthermore, as Winsler et al. (2002, p. 306) pointed out, "most of the (extant) research exploring the effects of age-composition on children's peer interactions has been conducted in laboratory or quasi-laboratory settings." In other words, with few exceptions, the extant literature lacks empirical research on preschool classroom age composition conducted in preschool classrooms. The exceptions (Bailey et al., 1993; Goldman, 1981; Winsler et al., 2002) have been limited to research conducted with a relatively small number of classrooms, typically restricted to a single organization. For example, Bailey et al.'s (1993) investigation tracked the developmental trajectories of just 59 children between the ages of 21 and 67 months. Goldman's (1981) sample consisted of just three



classrooms, and Winsler et al.'s (2002) sample included only four. Furthermore, the samples included in the vast majority of studies reviewed above came from affluent or middle-class suburban communities. From the perspective of triage at the level of public policy, an argument could be made for focusing attention first on the neediest and most underserved populations. Yet, urban samples of lower socioeconomic standing have typically been ignored in this literature. An additional limitation of the extant literature is that researchers have typically focused on a limited range of outcomes, perhaps masking a richer, more ambiguous phenomenon.

We undertook an investigation into the question of preschool classroom age composition on a scale far larger than previously conducted (806 children in 70 classrooms), a sample that included every preschool classroom under the administrative umbrella of a public school district in a mid-sized city (approximate population = 212,000). We included as outcomes in our research well-validated measures of various aspects of development (cognitive, social, and motor) and controlled for baseline levels on entering the classroom. Furthermore, we assessed and controlled for general classroom quality. As a result, this research should offer fresh and rich insight into questions concerning the consequences of preschool classroom age composition that can be extended more easily to urban populations.

## Method

### *Participants*

Participants included 806 preschool children from 70 preschool classrooms. The classrooms in the sample all fell under the administrative umbrella of a public school district in a mid-sized city in the northeastern United States. The mean age of children in the sample at Time 1 was 4.15 years ( $SD = 0.50$ ; 51% male). The ethnic composition included 57% African American, 17% White, 15% Hispanic, 2% Asian, and 9% children of unreported race or ethnicity. Overall, the students enrolled in the district were low in socioeconomic status (SES). Across the district, 86% of students came from homes below the poverty line, and 84% of the students qualified for free or reduced-price lunch. The mean number of students in a classroom was 16.34 (range = 8 to 21). The mean difference in age between the oldest and youngest child in a classroom was 1.58 years (range = 0.73 to 2.73). The mean standard deviation of children's age in classrooms was 0.44 (range = 0.10 to 0.70).

### *Measures*

#### *Child Observation Record (COR)*

This instrument (High/Scope Educational Research Foundation, 1992) measures several relevant domains of preschool children's development. The COR includes 32 items, with response options ranging from *least developed* to *most developed* on a 5-point scale, to be filled out by teachers immersed in the classroom environment from observations of naturally occurring behaviors. Before teachers used the COR they received 3 hr of training familiarizing them with this instrument. A recent investigation, focused on validating the COR, found that a three-factor solution, which included Cognitive Skills, Social Engagement, and Coordinated Movement/Motor Skills, best accounted for the correlations among the 32

COR items (Fantuzzo, Hightower, Grim, & Montes, 2002). Thus, in the present investigation, we used this three-factor solution, as well as the total COR (a combined score across all three subscales). Convergent and divergent validity studies have found significant relations between COR dimensions and other classroom competency constructs, such as peer play interactions (Penn Interactive Peer Play Scale; Fantuzzo, Coolahan, Mendez, McDermott, & Sutton-Smith, 1998; Fantuzzo et al., 2002; McWayne, Fantuzzo, & McDermott, 2004; Sekino & Fantuzzo, 2005), psychological adjustment (Teacher-Child Rating Scales [Fantuzzo et al., 2002; Perkins & Hightower, 1999, 2000] and Adjustment Scales for Preschool Intervention [Lutz, 1999; Lutz, Fantuzzo, & McDermott, 2002; Sekino & Fantuzzo, 2005]), receptive vocabulary (Peabody Picture Vocabulary Test-III; Dunn & Dunn, 1997; Fantuzzo et al., 2002), general learning skills (Preschool Learning Behaviors Scale [McDermott, Green, Francis, & Stott, 1996; McWayne et al., 2004; Sekino & Fantuzzo, 2005] and McCarthy Scale of Children's Abilities [McCarthy, 1972; Schweinhart, McNair, Barnes, & Lerner, 1993]), early reading ability (Test of Early Reading Ability, 3rd ed.; Reid, Hreski, & Hammill, 2001; Sekino & Fantuzzo, 2005), early math ability (Test of Early Mathematics Ability, 2nd ed.; Ginsburg & Broody, 1990; Sekino & Fantuzzo, 2005), kindergarten literacy (Dynamic Indicators of Basic Early Literacy Skills; Good & Kaminski, 2002; Sekino & Fantuzzo, 2005), and kindergarten academic success (Early Screening Inventory—Revised Kindergarten Version; Meisels, Marsden, Wiske, & Henderson, 1997; McWayne et al., 2004).

#### *Early Childhood Environment Rating Scale—Revised (ECERS-R)*

The ECERS-R is among the most widely used observational tools allowing for objective assessment of preschool classroom quality (Henry et al., 2004; Howes & Smith, 1995; Scarr, Eisenberg, & Deater-Deckard, 1994). The seven areas of classroom quality that the ECERS-R measures are space and furnishing, personal care routines, language and reasoning, activities, interaction, program structure, and parents and staff. Each area contains 5–10 items that represent various elements of that area. For the present investigation, we averaged scores across all seven areas to create an overall, composite measure of classroom quality.

### *Procedure*

Seventy teachers completed the COR for each child in their classroom, once in the fall and again in the spring. Teachers participated as part of an ongoing community assessment partnership that was open to all providers of services to preschool children in the area. Teachers were trained on the COR according to High/Scope standards, including procedures for observation and note taking, before completion of COR protocols. Trainees matched observations of videotaped classrooms against a standard to establish reliability levels of at least 80%. Master teachers and consultants were available to any teacher who had any questions or issues after the initial training for the entire school year. After COR ratings were completed, the first round of data sheets were returned to us for processing in October; the second round of data sheets were returned to us for processing in late April. Thus, the

typical length of time between COR assessments was 6–7 months. COR ratings were returned for 95% of the original sample.<sup>2</sup>

ECERS–R observations of classroom quality were conducted by 24 observers midway through the academic year (in the months of February, March, and April). For classroom observers in their 1st year of training, ECERS–R observers attended a 15-hr training program and reached an interrater reliability of 85% agreement with a master observer who was trained by the ECERS–R authors. For observers in their 2nd year, an additional 4–5 hr of training were required. All observers maintained an interrater reliability of 80% agreement, with 20% of their observations being checked.

## Results

### COR Scoring

We calculated COR scores for each of the three subscales—Cognitive Skills (Time 1 [T1],  $\alpha = .92$ ; Time 2 [T2],  $\alpha = .91$ ), Social Engagement (T1,  $\alpha = .93$ ; T2,  $\alpha = .93$ ), and Coordinated Movement/Motor (T1,  $\alpha = .87$ ; T2,  $\alpha = .90$ )—and the Combined/Total COR (T1,  $\alpha = .96$ ; T2,  $\alpha = .96$ ). The range for each COR subscale was from 0 to 5. Mean COR scores and standard deviations at Time 1 were, for Cognitive Skills, 2.08 and 0.69, respectively; for Social Engagement, 2.63 and 0.75, respectively; for Coordinated Movement/Motor, 2.68 and 0.68, respectively; and for Total COR, 2.43 and 0.66, respectively. Mean COR scores and standard deviations at Time 2 were, for Cognitive Skills, 3.07 and 0.77, respectively; for Social Engagement, 3.72 and 0.80, respectively; for Coordinated Movement/Motor, 3.65 and 0.80, respectively; and for Total COR, 3.48 and 0.72, respectively. Intercorrelations among the three COR factors ranged from .78 to .80 at Time 1 and from .71 to .76 at Time 2 (see Table 1). The high degree of correlation among the three COR factors indicates that analyzing each factor and the total COR score is justified.

### Primary Analysis

We used hierarchical linear modeling (HLM; Bryk & Raudenbush, 1992) for our primary tests. HLM appropriately addresses the hierarchically nested design of our data set, in which lower level units, children, were nested within a higher level unit, classroom. HLM treats classroom as a random rather than a fixed effect, thereby permitting generalizations of the findings to a wider population. Unlike regression models, this analysis allows for the possibility that the within-group (i.e., classroom) slopes may differ significantly from one classroom to another. Such an approach

makes it possible to see whether the relationship between child's age and change in COR score differs from classroom to classroom. For example, a positive main effect for classroom age range would imply that classrooms that are more mixed in terms of age composition are related to greater developmental growth, as assessed by COR change.

We calculated intraclass correlation coefficients (ICCs) for each of the four difference outcomes variables from the unconditional models: total COR score,  $ICC = .25$ ,  $\chi^2(69, N = 70) = 329.30$ ,  $p < .001$ ; COR Social,  $ICC = .25$ ,  $\chi^2(69, N = 70) = 330.56$ ,  $p < .001$ ; COR Motor,  $ICC = .23$ ,  $\chi^2(69, N = 70) = 301.19$ ,  $p < .001$ ; and COR Cognitive,  $ICC = .25$ ,  $\chi^2(69, N = 70) = 333.65$ ,  $p < .001$ . These findings confirmed that a HLM approach would provide substantial benefits over a standard fixed-effects model approach for the analysis of these data.

### Models 1–8: Chronological Age Effects

In the first set of models run, the primary classroom-level (Level 2) predictor was classroom chronological age composition, which was operationalized in two ways: (a) the range in chronological age between the youngest and oldest child in a given classroom and (b) the standard deviation of chronological age within a given classroom. Covariate predictors entered at the classroom level included the number of students in the classroom ( $M = 16.34$ ). At the person or child level (Level 1), child's chronological age at Time 1, gender, and COR assessment score at Time 1 were included as predictors. HLM estimates classroom-level and child-level effects simultaneously. Thus, classroom-level effects are statistically independent of child-level effects. All predictors (both classroom and child level) were centered on the sample (grand) means. The multi-level equations used for this basic model are described in Appendix A. This basic model was run independently using two different indicators of classroom chronological age composition (chronological age range in classroom and standard deviation of chronological ages in classroom) and four different outcomes (total COR score and each subscale from the COR), resulting in Models 1–8.

*Chronological age and the composite total COR.* The first version run of the basic model described above included chronological age range as the index of classroom chronological age composition and total COR score at Time 2 as the outcome measure (Model 1). The results of this hierarchical linear model are summarized in Table 2. Overall, there was a significantly negative main effect at the classroom level for classroom chronological age composition ( $\gamma_{01} = -.19$ ,  $p < .05$ ), and a nonsignificant effect for class size ( $\gamma_{02} = .02$ ,  $p = .37$ ). As expected, there were significant main effects observed at the child level for child's chronological age ( $\gamma_{10} = .23$ ,  $p < .001$ ) and Time 1 total COR score ( $\gamma_{20} = .83$ ,  $p < .001$ ). There was also a significant main effect at the child level for gender ( $\gamma_{30} = .06$ ,  $p < .05$ ), such that girls scored higher than boys on the COR at Time 2, controlling for

Table 1  
Correlations Between the Three COR Factors at Time 1 and Time 2 ( $N = 806$ )

Factor	1	2	3	4	5	6
1. T1 Social	—					
2. T1 Cognitive	.78**	—				
3. T1 Motor	.79**	.80**	—			
4. T2 Social	.60**	.50**	.52**	—		
5. T2 Cognitive	.57**	.63**	.55**	.71**	—	
6. T2 Motor	.52**	.52**	.58**	.75**	.76**	—

\*\* $p < .01$ .

<sup>2</sup> Forty-two children were dropped from the study (roughly 5%) because Time 2 COR scores were not available. These 42 children did not differ significantly from the retained sample ( $n = 806$ ) in terms of mean age, racial, or gender distribution.



Table 2  
Fixed Effects for Models 1–4 (Classroom Age Range as Focal Level 2 Predictor)

Parameter	Model 1: Total COR		Model 2: COR social		Model 3: COR motor		Model 4: COR cognitive	
	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE
Average classroom intercept ( $\gamma_{00}$ )	3.53***	0.04	3.75***	0.05	3.67***	0.04	3.12***	0.04
Age range ( $\gamma_{01}$ )	−0.19*	0.10	−0.22†	0.11	−0.16†	0.09	−0.19*	0.09
Class size ( $\gamma_{02}$ )	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02
Intercept for child's age ( $\gamma_{10}$ )	0.23***	0.05	0.20***	0.05	0.31***	0.05	0.33***	0.06
Age range × Child's age ( $\gamma_{11}$ )	−0.10	0.08	−0.04	0.09	−0.08	0.10	−0.12	0.11
Intercept for T1 total COR ( $\gamma_{20}$ )	0.83***	0.04	0.72***	0.04	0.73***	0.05	0.76***	0.05
Intercept for child's gender ( $\gamma_{30}$ )	0.06*	0.03	0.09**	0.03	0.09*	0.04*	0.03	0.03

Note. Model 1 = Total COR as dependent outcome; Model 2 = COR Social as dependent outcome; Model 3 = COR Motor as dependent outcome; Model 4 = COR Cognitive as dependent outcome; COR = Child Observation Record; Coeff = coefficient; T1 = Time 1.

\* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ . † $p < .10$ .

Time 1 score. Finally, the cross-level interaction between classroom chronological age composition and child's chronological age was nonsignificant ( $\gamma_{11} = -.10, p = .23$ ). This nonsignificant cross-level interaction suggested that the negative classroom-level main effect for chronological age composition was not moderated by the child-level chronological age of children in these classrooms.

The total variance at Level 1 (total COR at Time 2) was 1,648.63. This was estimated by specifying an unconditional model that included only a random intercept for each classroom. The error variance unaccounted for by the conditional Model 1 was reduced to 984.81. Comparison of these results indicates that 40.26% of the variance in total COR at Time 2 was accounted for by Model 1.

*Chronological age and the three COR subscales: Social, Motor, and Cognitive.* Next, we ran three hierarchical linear models using age range as the index of classroom age composition, independently predicting the three subscales from the COR as outcomes: Social, Motor, and Cognitive (Models 2, 3, and 4, respectively). The estimates from each model are summarized in Table 2, and the observed pattern of results was remarkably consistent across all three COR subscales. The negative main effect for classroom chronological age range was marginally significant when predicting COR Social and COR Motor and significant when predicting COR Cognitive. Child's chronological age was a significant predictor of all three COR subscales, as were Time 1 scores (for each subscale, respectively). Child's gender was a significant predictor of COR Social and COR Motor (in each case, girls scored higher), but did not significantly predict COR Cognitive. The cross-level interaction between child's chronological age and classroom chronological age range was nonsignificant for all three COR subscales.

The total variance at Level 1 was estimated by specifying an unconditional model that included only a random intercept for each classroom for each of the three COR subscales: Social, 1,813.15; Motor, 1,825.91; and Cognitive, 1,755.37. The error variance unaccounted for by Model 2 was reduced to 1,291.08; thus, the model accounted for 28.79% of the variance. The error variance unaccounted for by Model 3 was reduced to 1,367.06; thus, the model accounted for 25.13% of the variance. The error variance unaccounted for by Model 4 was reduced to 1,179.26; thus, the model accounted for 32.82% of the variance.

Next, we reran the four models described above, substituting the standard deviation of chronological ages for chronological age range at the classroom level (Models 5–8). Standard deviation of chronological ages in classrooms represents an alternative index of classroom chronological age composition that is more resistant to extreme values (i.e., if a single child were much younger or older than all the other children in a classroom). The pattern of results replicated those found using chronological age range as the index of classroom chronological age composition. For a summary of these results, see Table 3.

The error variance unaccounted for by Model 5 was reduced from 1,648.63 to 978.51; thus, the model accounted for 40.65% of the variance. The error variance unaccounted for by Model 6 was reduced from 1,813.15 to 1,284.49; thus, the model accounted for 29.16% of the variance. The error variance unaccounted for by Model 7 was reduced from 1,825.91 to 1,361.85; thus, the model accounted for 25.42% of the variance. The error variance unaccounted for by Model 8 was reduced from 1,755.37 to 1,172.09; thus, the model accounted for 33.23% of the variance.

### Models 9–16: Developmental Age Effects

In the second set of models run, the primary classroom-level (Level 2) predictor was classroom developmental age composition. Developmental age was operationalized as COR score at Time 1. As such, in the interest of conceptual clarity, in the following models Time 1 COR score is referred to as *developmental age*. Developmental age range was operationalized in two ways: the range in score between the lowest and highest Time 1 COR score in a given classroom and the standard deviation of COR scores at Time 1 within a given classroom. Covariate predictors entered at the classroom level included the number of students in the classroom ( $M = 16.34$ ).

At the person or child level (Level 1), child's developmental age (i.e., Time 1 COR score), gender, and age at Time 1 were included as predictors. Finally, we included a term representing the cross-level interaction between child's developmental age (Time 1 COR) and classroom developmental age composition (range in COR scores or standard deviation of COR scores). HLM estimates classroom-level and child-level effects simultaneously. Thus, classroom-level effects are statistically independent of child-level

Table 3

*Fixed Effects for Models 5–8 (Classroom Age Standard Deviation as Focal Level 2 Predictor)*

Parameter	Model 5: Total COR		Model 6: COR social		Model 7: COR motor		Model 8: COR cognitive	
	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE
Average classroom intercept ( $\gamma_{00}$ )	3.53***	0.04	3.75***	0.05	3.67***	0.04	3.12***	0.04
Age SD ( $\gamma_{01}$ )	−0.74*	0.31	−0.79*	0.34	−0.60*	0.30	−0.71*	0.30
Class size ( $\gamma_{02}$ )	0.01	0.02	0.01	0.02	0.02	0.02	0.01	0.02
Intercept for child's age ( $\gamma_{10}$ )	0.20***	0.05	0.17***	0.05	0.20***	0.05	0.30***	0.06
Age SD $\times$ Child's age ( $\gamma_{11}$ )	0.17	0.32	0.44	0.34	−0.23	0.45	0.18	0.43
Intercept for T1 Total COR ( $\gamma_{20}$ )	0.82***	0.04	0.72***	0.04	0.73***	0.05	0.76***	0.04
Intercept for child's gender ( $\gamma_{30}$ )	0.06*	0.03	0.09**	0.03	0.09*	0.04	0.03	0.03

Note. Model 1 = Total COR as dependent outcome; Model 2 = COR Social as dependent outcome; Model 3 = COR Motor as dependent outcome; Model 4 = COR Cognitive as dependent outcome; COR = Child Observation Record; Coeff = coefficient; T1 = Time 1.

\* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

effects. All predictors (both classroom and child level) were centered on the sample (grand) means. The multilevel equations used for the second basic model are described in Appendix B. (As noted earlier, the multilevel equations used for the first basic model are described in Appendix A.) Paralleling Models 1–8, the second basic model was run independently, using two different indicators of classroom developmental age composition (developmental age range in classroom and standard deviation of developmental ages in classroom) and four different outcomes (total COR score and each subscale from the COR), resulting in Models 9–16.

*Developmental age and the composite COR total.* The basic model described above was first run using developmental age range as the index of classroom developmental age composition and total COR score at Time 2 as the outcome measure (Model 9). The results of this hierarchical linear model are summarized in Table 4. Overall, there was a significantly negative main effect at the classroom level for classroom developmental age range ( $\gamma_{01} = -.23$ ,  $p < .01$ ). The classroom-level effect for class size was nonsignificant ( $\gamma_{02} = .02$ ,  $p = .31$ ). As expected, there were significant main effects observed at the child level for child's developmental age ( $\gamma_{10} = .86$ ,  $p < .001$ ) and child's chronological age ( $\gamma_{20} = .22$ ,  $p < .001$ ). There was also a

significant child-level main effect for gender ( $\gamma_{30} = .06$ ,  $p < .05$ ). Finally, a significant cross-level interaction between classroom developmental age range and child's developmental age was observed ( $\gamma_{11} = -.17$ ,  $p < .01$ ). This cross-level interaction suggested that there was little difference in terms of Time 2 COR for children lower in developmental age (i.e., 1 standard deviation below the grand mean Time 1 Total COR score) in classrooms with mixed versus restricted developmental age ranges. However, for children higher in developmental age (i.e., 1 standard deviation above the grand mean Time 1 Total COR score), there was a significant advantage to being in a classroom with a restricted developmental age range versus a classroom with a wide range in developmental age. This interaction is illustrated in Figure 1.

The total variance at Level 1 (total COR at Time 2) was 1,648.63. This was estimated by specifying an unconditional model that included only a random intercept for each classroom. The error variance unaccounted for by the conditional Model 1 was reduced from 1,648.63 to 989.45; thus, the model accounted for 40.00% of the variance.

*Developmental age and the three COR subscales: Social, Motor, and Cognitive.* Next, three hierarchical linear models using developmental age range as the index of classroom developmental

Table 4

*Fixed Effects for Models 9–12 (Classroom Developmental Age Range as Focal Level 2 Predictor)*

Parameter	Model 9: Total COR		Model 10: COR social		Model 11: COR motor		Model 12: COR cognitive	
	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE
Average classroom intercept ( $\gamma_{00}$ )	3.54***	0.04	3.75***	0.05	3.68***	0.04	3.13***	0.04
Developmental age range ( $\gamma_{01}$ )	−0.23**	0.07	−0.21*	0.09	−0.15*	0.07	−0.23**	0.07
Class size ( $\gamma_{02}$ )	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Intercept for child's age ( $\gamma_{10}$ )	0.86***	0.05	0.75***	0.04	0.74***	0.05	0.78***	0.05
Developmental age range $\times$ Child's developmental age ( $\gamma_{11}$ )	−0.17**	0.05	−0.10*	0.05	−0.08	0.07	−0.08	0.06
Intercept for child's age ( $\gamma_{20}$ )	0.22***	0.04	0.20***	0.05	0.30***	0.05	0.32***	0.06
Intercept for child's gender ( $\gamma_{30}$ )	0.06*	0.03	0.09**	0.03	0.09*	0.04	0.03	0.03

Note. Model 1 = Total COR as dependent outcome; Model 2 = COR Social as dependent outcome; Model 3 = COR Motor as dependent outcome; Model 4 = COR Cognitive as dependent outcome; COR = Child Observation Record; Coeff = coefficient.

\* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .



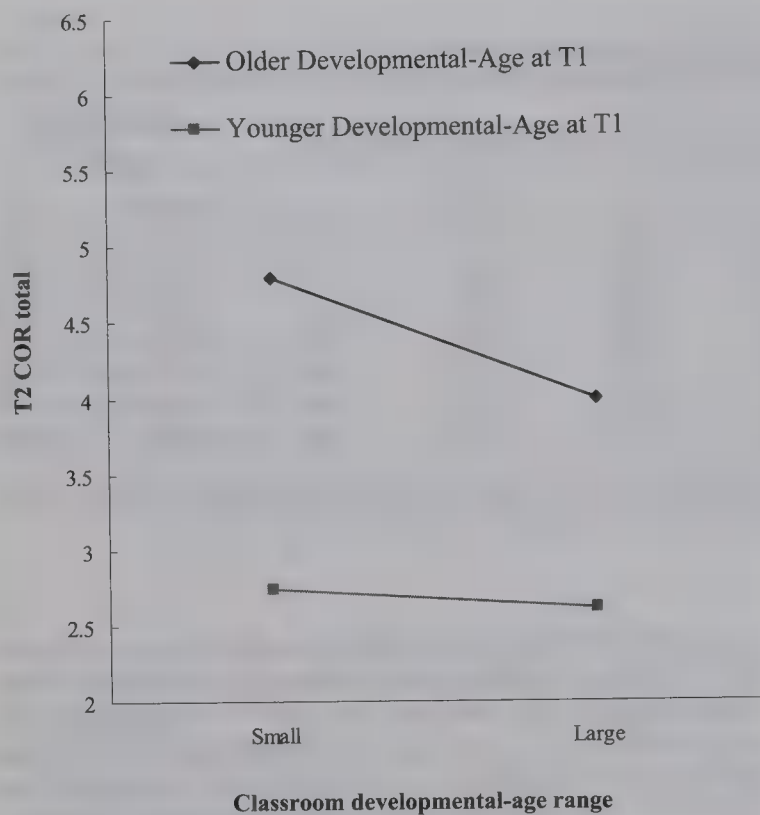


Figure 1. Classroom developmental age range by child's developmental age predicting total Child Observation Record (COR) score at Time 2, controlling for total COR score at Time 1.

age composition were run independently, predicting the three COR subscales as outcomes: Social, Motor, and Cognitive (Models 10, 11, and 12, respectively). The estimates from each model are summarized in Table 4, and the observed pattern of results was remarkably consistent across all three COR subscales. Classroom-level developmental age range was a significantly negative predictor of all three subscales from the COR. Class size was not a significant predictor of any of the three COR subscales. Child's chronological age and developmental age significantly predicted all three COR subscales. Child's gender significantly predicted COR Social and COR Motor (girls scored higher), but not COR Cognitive. Finally, the cross-level interaction between classroom-level developmental age range and child's developmental age was significant only with regard to COR Social, but did not significantly predict COR Motor or COR Cognitive. These results are summarized in Table 4.

Next, we reran the four models described above, substituting the standard deviation of developmental ages as the indicator of developmental age composition at the classroom level (Models 13–16). Standard deviation of developmental ages in classrooms represents an alternative index of classroom developmental age composition that is more resistant to extreme values (i.e., if a single child scored much higher or lower on the Time 1 COR relative to all the other children in a given classroom). The pattern of results replicated those found using developmental age range as the index of classroom developmental age composition. These results are summarized in Table 5.

**Classroom quality.** We next included classroom quality (ECERS-R) in the models described above exploring the influence of both chronological (Models 1–8) and developmental (Models

9–16) age composition on residual COR change. The significant effects reported for classroom age composition all remained significant when controlling for classroom quality. That is, classroom-level variance in age, chronological or developmental, was negatively related to residual COR change. Furthermore, there were no significant main effects found for classroom quality and no significant classroom-level interactions between classroom quality and classroom age composition.

It is, however, worth noting that the ECERS-R scores for the classrooms in this sample were uniformly approaching ceiling levels for excellence. The maximum possible score for overall classroom quality on this scale is 7.0. In the school district sampled, the mean score for overall classroom quality was 6.42 ( $Mdn = 6.75$ ), with a standard deviation of only 0.77. Thus, the level of ECERS-R-rated classroom quality in this sample was greater than 2 standard deviations above the national average (approximately 5.0), as previously reported by Montes, Hightower, Brugger, and Moustafa (2005).

**SES.** Next, an indicator of classroom-level SES was included in the models (1–16) described above. In the absence of direct indicators of family income, we used mothers' level of education as an indirect indicator of child-level SES. Studies that have explored the relation between level of education and income typically report strong correlations (e.g.,  $r = .51$ ; Robert, 1998); researchers often use these variables as interchangeable indicators of SES in multivariate models owing to the high degree of shared variance (e.g., Winkleby, Kraemer, Ahn, & Varady, 1998). Thus, mean and median mothers' highest completed level of education were used as indicators of SES at the classroom level.<sup>3</sup> Across all 70 classrooms in the sample, the mean level of mothers' education was 3.25 ( $SD = 1.09$ ), corresponding to a level of education between high school diploma or GED and some college. The significant effects reported for classroom age composition all remained significant when controlling for classroom-level SES. That is, greater variance in age, chronological or developmental, at the classroom level was negatively related to residual COR change. Furthermore, no significant main effects were found for classroom SES, and there were no significant classroom-level interactions between classroom SES and classroom age composition.

## General Discussion

This investigation represents a unique and important contribution to the literature on preschool classroom age composition in a number of respects. First, the study included a sample far larger than that in any previously conducted research, 806 children from 70 different preschool classrooms. Second, this research is among the first to use a well-validated assessment of early childhood development (i.e., the COR) in a variety of domains (social, motor, and cognitive) and to include assessments at two time points (spaced approximately 6 months apart). Thus, we were able to explore the influence of classroom age composition on residualized change in various aspects of development. Additionally, important covariates, such as general classroom quality (i.e., the

<sup>3</sup> Mother's education was available for 88% of the children in the sample. To calculate mean and median mothers' education at the classroom level, missing values were replaced with the classroom (or group) mean.

Table 5  
Fixed Effects for Models 13–16 (Classroom Developmental Age Standard Deviation as Focal Level 2 Predictor)

Parameter	Model 13: Total COR		Model 14: COR social		Model 15: COR motor		Model 16: COR cognitive	
	Coeff	SE	Coeff	SE	Coeff	SE	Coeff	SE
Average classroom intercept ( $\gamma_{00}$ )	3.53***	0.04	3.75***	0.05	3.67***	0.04	3.12***	0.04
Developmental age SD ( $\gamma_{01}$ )	-0.71**	0.23	-0.59*	0.28	-0.46*	0.21	-0.72**	0.21
Class size ( $\gamma_{02}$ )	0.01	0.02	0.01	0.02	0.02	0.02	0.01	0.02
Intercept for child's age ( $\gamma_{10}$ )	0.86***	0.05	0.75***	0.04	0.73***	0.05	0.78***	0.05
Developmental age SD $\times$ Child's developmental age ( $\gamma_{11}$ )	-0.48*	0.19	-.29 <sup>†</sup>	0.15	-0.07	0.19	-0.22	0.22
Intercept for child's age ( $\gamma_{20}$ )	0.22***	0.04	0.20***	0.05	0.30***	0.05	0.32***	0.06
Intercept for child's gender ( $\gamma_{30}$ )	0.06	0.03	0.09**	0.03	0.09*	0.04	0.03	0.03

Note. Model 1 = Total COR as dependent outcome; Model 2 = COR Social as dependent outcome; Model 3 = COR Motor as dependent outcome; Model 4 = COR Cognitive as dependent outcome; COR = Child Observation Record; Coeff = coefficient.

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ . <sup>†</sup>  $p < .10$ .

ECERS-R) and SES, were included. Finally, this research focuses on a population (i.e., urban preschool children) that has been understudied, yet is extremely important from the perspective of policymakers and others interested in social welfare.

The results were surprising, insofar as they contradicted assertions made by advocates of mixed-age classrooms (Bredekamp, 1987; Katz et al., 1990), yet remarkably consistent across indicators of classroom age composition and domains of development. We consistently observed a significant main effect at the classroom level for classroom age composition, which suggested that a wide range in children's ages within a classroom (and high standard deviations in terms of age) was negatively related to development. This was true in terms of overall development (i.e., total COR score) and, independently, in terms of social, motor, and cognitive domains of development (COR subscales).

Furthermore, the negative relation observed between classroom (chronological) age composition and developmental outcomes was replicated using an indicator of classroom developmental age composition (range and standard deviation of Time 1 COR scores). This alternative approach to exploring the impact of classroom age composition was important for a number of reasons.

First, the parallel set of findings observed for classroom developmental age range helps to mitigate potential confounds that might have resulted from nonrandom assignment to classrooms. For example, one might imagine that the oldest children in mixed-age classrooms may be more likely to have been held back and that the brightest children are culled from these mixed-age classrooms when they become older. Those classrooms with wider distributions in age might also have a higher frequency of older children who are slow learners or are developmentally delayed. However, the models run using Time 1 COR scores as an indicator of developmental age help to address and dismiss the notion that older children in mixed-age classrooms in our sample were developmentally delayed or otherwise noncomparable to older children in single-age classrooms, as chronological age was separated from developmental age, and the pattern of results did not change.

Second, the fact that a wider distribution of both chronological and developmental age in preschool classrooms was negatively related to development, coupled with a large and diverse sample of classrooms, makes a strong case for generalizing this effect to other urban populations.

Third, the use of Time 1 COR scores as an indicator of developmental age allowed us to identify a cross-level interaction between classroom developmental age composition and child's developmental age, which further specifies the nature of the significant main effect found for classroom developmental age composition. This cross-level interaction suggested that children high in developmental age were most negatively influenced by assignment to classrooms with wider distributions of developmental age, whereas younger children did not seem to be significantly affected by classroom assignment.

As noted, these findings stand in contrast to enthusiasm that has been expressed for movement toward more mixed-age classrooms in preschool education (e.g., Bredekamp, 1987; Katz et al., 1990). As we have argued, the empirical evidence for this enthusiasm has been decidedly mixed and inconclusive to this point and, more important, has relied on relatively small samples, typically in suburban settings. In this context, the present research strongly suggests that reconsideration of the issue of classroom age composition in early childhood education is warranted.

Most relevant, this research speaks to the consequences of variance in classroom age composition as it naturalistically occurs in urban preschool classrooms today. Although it could well be that mixed-age contexts have the potential to benefit children, in this large and diverse urban sample this was not the case. The opposite proved true. For reasons yet to be fully elaborated, and which merit further research, older children in mixed-age classrooms developed at a slower rate than older children in single-age classrooms.

### Theoretical Implications

The findings from the present investigation strongly support the theory-based predictions offered by Piaget (1932) and others, who argued that interacting with peers who are close in age and ability will result in optimal learning. At the same time, these findings are not entirely inconsistent with predictions offered by Vygotsky (1930/1978) and others, who argued for mixed-age interaction principally on the basis of the merits implicit for younger children in these contexts. These authors posited that younger children could benefit from interacting with older, more competent children. In this regard, their prediction was unsupported. However,



implicit in the Vygotskian model are the potential negative consequences for those who are older and more competent in mixed-age interaction. Thus, in terms of the relative advantage or disadvantage for different children in mixed-age interaction, the data from this study provide some support for the assertions made by Vygotsky and his colleagues. That is, one interpretation of the Vygotskian model is that variance in age composition or ability among interaction partners should be relatively less beneficial for those older children who are higher in ability. Our investigation supported this relativistic aspect of the Vygotskian model, such that although the influence of developmental age variance was negligible for developmentally younger children, there were significant negative consequences for developmentally older children.

The theoretical accounts for the influence of mixed- versus restricted-age groupings offered by Piaget (1932) and Vygotsky (1930/1978) focused primarily on the influence of peer interaction. However, it is likely that those supervising children (teachers, in the present investigation) may behave differently in mixed- versus restricted-age groupings as well. That is, as a result of being placed in mixed-age versus restricted-age classrooms, teachers may change their modes of instruction and interaction. Although exploring the dynamics through which teachers may be influenced and, in turn, exert influence on the children in mixed- versus restricted-age classrooms was beyond the scope of this study, it remains an important issue to consider.

### *Limitations and Conclusions*

Differences in these findings relative to other published studies may, in part, be attributed to the time horizon used in assessing the influence of classroom age composition. The vast majority of studies on classroom age composition have been cross-sectional (i.e., single time point as opposed to change) in design (e.g., Dunn et al., 1996; Field, 1982; Goldman, 1981; Urberg & Kaplan, 1986; for exceptions, see Bailey et al., 1993, and Winsler et al., 2002). A strength of the present investigation was that development was assessed longitudinally. However, the full influence of mixed-age classrooms may take several years to manifest. One future direction for this line of research would be to explore the influence of classroom age composition on developmental change across multiple time points, extending from 6 months (as in the present research) to several years after initial exposure to these classrooms.

A strength of this research was the use of assessments of constructs at multiple time points; however, data were only collected at two time points. A logical extension of the basic approach we have outlined in this research would involve assessments at three or more time points. This would permit exploration into the possibility of both linear and nonlinear trajectories.

Another limitation of this research involves the correlational nature of these data. Empirical investigations that manipulate the age composition of preschool classrooms, with random assignment to condition, are warranted. As an alternative, when large-scale institutional policy changes are made affecting the age composition of preschool classrooms, such cases represent pseudoexperiments that may be used to more strongly establish a causal relation between classroom age composition and developmental outcomes (e.g., Winsler et al., 2002).

Finally, the uniformly high ratings of classroom quality (ECERS-R) in this sample were not conducive to appropriately

testing the potential for classroom quality to moderate the negative effect of age variance observed in mixed-age classrooms. Further investigations are called for, targeting samples that include classrooms with greater variance in quality and the inclusion of additional indicators of classroom quality.

Despite these limitations, this investigation should be considered a significant contribution, as it represents an empirical investigation of preschool classroom age composition on a far larger scale than previously completed. It introduces a novel, and potentially generative, methodological approach for investigating the influence of classroom age composition (i.e., treating age composition as a continuous variable). Furthermore, this question was explored in terms of both chronological and developmental age composition in these classrooms, and the results were remarkably consistent across indicators of classroom composition and domains of development. The results support initiatives to maintain and facilitate movement toward preschool classrooms with more restricted age composition, at least insofar as age composition is a factor considered independently from curriculum considerations. As noted, the influence of age composition may be dynamically related to teachers' modes of instruction, to the extent that teachers are influenced by classroom age composition. The present investigation leaves open the possibility that mixed-age preschool classrooms coupled with curricula tailored to such climates could indeed be as good as or even superior to the typical restricted-age classroom. Specifically, our data suggest that such mixed-age-classroom-tailored curricula should address the needs of more mature, developmentally advanced children, who seem to have fared the worst under these conditions in our sample.

Ultimately, a thorough analysis of this issue must also take into consideration the economics of mixed- versus restricted-age classrooms. In areas of the country in which preschool classrooms are easily filled to capacity, it seems most prudent at this juncture to advocate for more restricted age composition. However, in parts of the country in which preschool classrooms are less easily filled, it may become significantly more cost effective to permit greater diversity in classroom age composition. As such, we cannot unilaterally condemn mixed-age classrooms but would instead implore educators compelled to work in these settings to make special efforts to provide individualized, optimal challenges for children who may otherwise be disadvantaged.

### *References*

- Bailey, D. B., Burchinal, M. R., & McWilliams, R. A. (1993). Age of peers and early childhood development. *Child Development*, 64, 848-862.
- Bandura, A. (1986). *Social foundations of thought and action*. Englewood Cliffs, NJ: Prentice Hall.
- Blasco, P. M., Bailey, D. B., & Burchinal, M. A. (1993). Dimensions of mastery in same-age and mixed-age integrated classrooms. *Early Childhood Research Quarterly*, 8, 193-206.
- Bredekamp, S. (Ed.). (1987). *Developmentally appropriate practice in early childhood programs serving children from birth through age 8: Expanded edition*. Washington, DC: National Association for the Education of Young Children.
- Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.
- Byrnes, D. A., Shuster, T., & Jones, M. (1994). Parent and student views on multiage classrooms. *Journal of Research in Childhood Education*, 9, 15-23.



- Chapman, M. L. (1994). Literacy learning in a primary multiage classroom: Some findings from an investigation of peer writing events. In C. K. Kinzer & D. J. Leu (Eds.), *Multidimensional aspects of literacy research, theory, and practice. Forty-third Yearbook of the National Reading Conference* (pp. 550–559). Chicago: The National Reading Conference.
- Derscheid, L. E. (1997). Mixed-age grouped preschoolers' moral behavior and understanding. *Journal of Research in Childhood Education, 11*, 147–151.
- Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Services.
- Dunn, L., Kontos, S., & Potter, L. (1996). Mixed-age interactions in family child care. *Early Education and Development, 7*, 349–366.
- Elliot, A. E., & Dweck, C. (2005). *Handbook of competence motivation*. New York: Plenum Press.
- Ellis, S., Rogoff, B., & Cromer, C. C. (1981). Age segregation in children's social interactions. *Developmental Psychology, 17*, 399–407.
- Fantuzzo, J., Coolahan, K. C., Mendez, J., McDermott, P. A., & Sutton-Smith, B. (1998). Contextually-relevant validation of constructs of peer play with African American Head Start children: Penn Interactive Peer Play Scale. *Early Childhood Research Quarterly, 13*, 411–431.
- Fantuzzo, J., Hightower, D., Grim, S., & Montes, G. (2002). Generalization of the Child Observation Record: A validation study for diverse samples of urban, low-income preschool children. *Early Childhood Research Quarterly, 17*, 106–125.
- Field, T. (1982). Same-sex preferences of preschool children: An artifact of same-age grouping? *Child Study Journal, 12*, 151–159.
- French, D. C., Waas, G. A., Stright, A. L., & Baker, J. A. (1986). Leadership asymmetries in mixed-age children's groups. *Child Development, 57*, 1277–1283.
- Gallagher, J. J. (1986). The need for programs for young gifted children. *Topics in Early Childhood Special Education, 6*, 1–8.
- Ginsburg, H. P., & Broody, A. J. (1990). *Test of Early Mathematics Ability* (2nd ed.). Austin, TX: PRO-ED.
- Goldman, J. A. (1981). Social participation of preschool children in same-versus mixed-age groups. *Child Development, 52*, 644–650.
- Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic indicators of basic early literacy skills* (6th ed.) Retrieved from University of Oregon, Institute for the Development of Education Achievement: <http://dibels.uoregon.edu>
- Guralnick, M. J. (1981). The social behavior of preschool children at different developmental levels: Effects of group composition. *Journal of Experimental Child Psychology, 31*, 115–130.
- Hartup, W. W. (1993). Peer relationships. In P. H. Mussen (Series Ed.) & E. M. Hetherington (Vol. Ed.), *Handbook of child psychology: Vol. 4. Socialization, personality, and social development* (4th ed., pp. 103–196). New York: Wiley.
- Henry, G. T., Ponder, B. D., Rickman, D. K., Mashburn, A. J., Henderson, L. W., & Gordon, C. S. (2004). *An evaluation of the implementation of Georgia's pre-k program: Report of the findings from the Georgia early childhood study (2002–2003)*. Retrieved November 24, 2006 from <http://aysps.gsu.edu/publications/2005/GAPrek2004.pdf>
- High/Scope Educational Research Foundation. (1992). *Manual: High/Scope Child Observation Record for ages 2.5–6*. Ypsilanti, MI: High/Scope Press.
- Howes, C., & Farver, J. (1987). Social pretend play in 2-year olds: Effects of age of partner. *Early Childhood Research Quarterly, 2*, 305–314.
- Howes, C., & Smith, E. W. (1995). Relations among child care quality, teacher behavior, children's play activities, emotional security, and cognitive activity in child care. *Early Childhood Research Quarterly, 10*, 381–404.
- Katz, L. G., Evangelou, D., & Hartman, J. A. (1990). *The case for mixed-age groupings in early childhood education*. Washington, DC: National Association for the Education of Young Children.
- Lloyd, L. (1999). Multi-age classes and high ability students. *Review of Education Research, 69*, 187–212.
- Lougee, M., Grueneich, R., & Hartup, W. (1977). Social interaction in same- and mixed-age dyads of preschool children. *Child Development, 48*, 1353–1361.
- Lutz, M. N. (1999). *Contextually relevant assessment of the emotional and behavioral adjustment of Head Start children*. Unpublished doctoral dissertation, University of Pennsylvania.
- Lutz, M. N., Fantuzzo, J., & McDermott, P. (2002). Contextually relevant assessment of the emotional and behavioral adjustment of low-income preschool children. *Early Childhood Research Quarterly, 17*, 338–355.
- McCarthy, D. A. (1972). *Manual for the McCarthy Scales of Children's Abilities*. San Antonio, TX: Psychological Corporation.
- McDermott, P. A., Green, L. F., Francis, J. M., & Stott, D. H. (1996). *Preschool Learning Behaviors Scales*. Philadelphia: Edumetric & Clinical Science.
- McWayne, C., Fantuzzo, J. W., & McDermott, P. A. (2004). Preschool competency in context: An investigation of the unique contribution of child competencies to early academic success. *Developmental Psychology, 40*, 633–645.
- Meisels, S. J., Marsden, D. B., Wiske, M. S., & Henderson, L. W. (1997). *The Early Screening Inventory* (Rev. ed.). Ann Arbor, MI: Rebus.
- Montes, G., Hightower, A. D., Brugger, L., & Moustafa, E. (2005). Quality child care and socio-emotional risk factors: No evidence of diminishing returns. *Early Childhood Research Quarterly, 20*, 361–372.
- National Association for the Education of Young Children. (1996). *Developmentally appropriate practice in early childhood programs serving children from birth through age 8* (NAEYC position statement) [Online]. Retrieved September 21, 2006 from <http://www.naeyc.org/about/positions/daptoc.asp>
- Nicholls, J. G. (1989). *The competitive ethos and democratic education*. Cambridge, MA: Harvard University Press.
- Perkins, P. E., & Hightower, A. D. (1999, April). *Construct validity of the Teacher-Child Rating Scale—Revised*. Poster presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.
- Perkins, P. E., & Hightower, A. D. (2000, August). *Improving the assessment of children's problem behaviors and competencies: Refinement of the Teacher-Child Rating Scale*. Presented at the 108th Annual Convention of the American Psychological Association, Anaheim, California.
- Piaget, J. (1932). *The moral judgment of the child*. London: Kegan Paul.
- Reid, D. K., Hreski, W. P., & Hammill, D. D. (2001). *Test of Early Reading Ability* (3rd ed.). Austin, TX: PRO-ED.
- Robert, S. A. (1998). Community-level socioeconomic status effects on adult health. *Journal of Health and Social Behavior, 39*, 18–37.
- Rogers, D. L., & Ross, D. D. (1986). Encouraging positive social interaction among young children. *Young Children, 41*, 12–17.
- Roopnarine, J. L., Ahmeduzzaman, M., Donnely, S., Gill, P., Mennis, A., Arky, L., et al. (1992). Social-cognitive play behaviors and playmate preferences in same-age and mixed-age classrooms over a 6-month period. *American Educational Research Journal, 29*, 757–776.
- Roopnarine, J. L., & Johnson, J. E. (1984). Socialization in a mixed-age experimental program. *Developmental Psychology, 20*, 828–832.
- Rothstein-Fisch, C., & Howes, C. (1988). Toddler peer interaction in mixed-age groups. *Journal of Applied Developmental Psychology, 9*, 211–218.
- Rubin, K. H., Bukowski, W., & Parker, J. G. (1998). Peer interactions, relationships, and groups. In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (5th ed., pp. 619–700). New York: Wiley.
- Rubin, K. H., Maioni, T. L., & Hornung, M. (1976). Free play behaviors in middle- and lower-class preschoolers: Parten and Piaget revisited. *Child Development, 47*, 414–419.



- Scarr, S., Eisenberg, M., & Deater-Deckard, K. (1994). Measurement of quality in child care centers. *Early Childhood Quarterly*, 9, 131–151.
- Schweinhart, L. J., McNair, S., Barnes, H., & Larner, M. (1993). Observing young children in action to assess their development: The High/Scope Child Observation Record study. *Educational and Psychological Measurement*, 53, 445–455.
- Sekino, Y., & Fantuzzo, J. (2005). Validity of the Child Observation Record: An investigation of the relationship between COR dimensions and social-emotional and cognitive outcomes for head start children. *Journal of Psychoeducational Assessment*, 23, 242–261.
- Tach, L. M., & Fargas, G. (2006). Learning-related behaviors, cognitive skills, and ability grouping when school begins. *Social Science Research*, 35, 1048–1079.
- Tudge, J. R. H. (1992). Processes and consequences of peer collaboration: A Vygotskian analysis. *Child Development*, 63, 1364–1379.
- Urberg, K. A., & Kaplan, M. G. (1986). Effects of classroom age-composition on the play and social behaviors of preschool children. *Journal of Applied Developmental Psychology*, 7, 403–415.
- Veenman, S. (1995). Cognitive and noncognitive effects of multigrade and multi-age classes: A best-evidence synthesis. *Review of Educational Research*, 65, 319–381.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher mental processes*. In M. Cole, V. John-Steiner, S. Scribner, & E. Souberman (Eds. & Trans.), Cambridge, MA: Harvard University Press. (Original work published in 1930).
- Whaley, K. L., & Kantor, R. (1992). Mixed-age groupings in infant/toddler child care: Enhancing developmental processes. *Child & Youth Care Forum*, 21, 369–384.
- Winkleby, M. A., Kraemer, H. C., Ahn, D. K., & Varady, A. N. (1998, July 22). Ethnic and socioeconomic differences in cardiovascular disease risk factors: Findings for women from the third national health and nutrition examination survey, 1988–1994. *Journal of the American Medical Association*, 280, 356–362.
- Winsler, A., Caverly, S. L., Willson-Quayle, A., Carlton, M. P., Howell, C., & Long, G. N. (2002). The social and behavioral ecology of mixed-age and same-age preschool classrooms: A natural experiment. *Applied Developmental Psychology*, 23, 305–330.

## Appendix A

### Multilevel Equations for Models 1–8

We estimated child-level (Level 1) Time 2 (T2) Child Observation Record (COR) scores by means of the following equation:

$$\begin{aligned} \text{T2 COR} = & \beta_0 + \beta_1(\text{Child's Age}) + \beta_2(\text{T1 COR}) \\ & + \beta_3(\text{Gender}) + e, \end{aligned}$$

where  $\beta_0$  refers to the intercept (i.e., the average COR score at Time 2);  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  represent the maximum likelihood estimate of the population slopes estimating Time 2 COR score from the child's age, child's T1 COR score, and gender, respectively; and  $e$  is error.

We estimated classroom-level (Level 2) effects as follows:

$$\beta_0 = \gamma_{00} + \gamma_{01}(\text{Age Composition}) + \gamma_{02}(\text{Class Size}) + u_0,$$

$$\beta_1 = \gamma_{10} + \gamma_{11}(\text{Age Composition}) + u_1,$$

$$\beta_2 = \gamma_{20} + u_2, \text{ and}$$

$$\beta_3 = \gamma_{30} + u_3.$$

We estimated the combined model as follows:

$$\begin{aligned} \text{T2 COR} = & \gamma_{00} + \gamma_{01}(\text{Age Composition}) \\ & + \gamma_{02}(\text{Class Size}) + \gamma_{10}(\text{Child's Age}) \end{aligned}$$

$$\begin{aligned} & + \gamma_{11}(\text{Age Composition})(\text{Child's Age}) \\ & + \gamma_{20}(\text{T1 COR}) + \gamma_{30}(\text{Gender}) + u_0 + u_1(\text{Child's Age}) \\ & + u_2(\text{T1 COR}) + u_3(\text{Gender}) + e, \end{aligned}$$

where  $\gamma_{00}$  refers to the child-level intercept for an average classroom;  $\gamma_{01}$  represents the maximum likelihood estimate of the population slope estimating average levels of Time 2 COR scores across all classrooms from classroom-level age composition;  $\gamma_{02}$  represents the maximum likelihood estimate of the population slope estimating average levels of Time 2 COR scores across all classrooms from classroom-level size (i.e., the number of students in the classroom) and  $u_0$  represents error in estimating this intercept;  $\gamma_{10}$  represents the maximum likelihood estimate of the population slopes estimating Time 2 COR score from child's age and  $u_1$  represents error related to that estimate;  $\gamma_{20}$  represents the maximum likelihood estimate of the population slopes estimating Time 2 COR score from Time 1 COR score and  $u_2$  represents error related to that estimate;  $\gamma_{30}$  represents the maximum likelihood estimate of the population slopes estimating Time 2 COR score from gender and  $u_3$  represents error related to that estimate; and  $\gamma_{11}$  represents the maximum likelihood estimate of the cross-level interaction between child-level age and classroom-level age composition.

## Appendix B

## Multilevel Equations for Models 9–16

We estimated child-level (Level 1) Time 2 COR scores by the following equation:

$$\begin{aligned} \text{T2 COR} = & \beta_0 + \beta_1(\text{Child's Developmental Age}) \\ & + \beta_2(\text{Child's Age}) + \beta_3(\text{Gender}) + e, \end{aligned}$$

where  $\beta_0$  refers to the intercept (i.e., the average COR score at Time 2);  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  represent the maximum likelihood estimate of the population slopes estimating Time 2 COR score from the child's developmental age, child's age, and gender, respectively; and  $e$  is error.

We estimated classroom-level (Level 2) effects as follows:

$$\begin{aligned} \beta_0 = & \gamma_{00} + \gamma_{01}(\text{Developmental Age Composition}) \\ & + \gamma_{02}(\text{Class Size}) + u_0, \\ \beta_1 = & \gamma_{10} + \gamma_{11}(\text{Developmental Age Composition}) + u_1, \\ \beta_2 = & \gamma_{20} + u_2, \text{ and} \\ \beta_3 = & \gamma_{30} + u_3. \end{aligned}$$

We estimated the combined model as follows:

$$\begin{aligned} \text{T2 COR} = & \gamma_{00} + \gamma_{01}(\text{Developmental Age Composition}) \\ & + \gamma_{02}(\text{Class Size}) + \gamma_{10}(\text{Child's Developmental Age}) \\ & + \gamma_{11}(\text{Developmental Age Composition}) \\ & (\text{Developmental Age}) + \gamma_{20}(\text{Child's Age}) \end{aligned}$$

$$\begin{aligned} & + \gamma_{30}(\text{Gender}) + u_0 + u_1(\text{Child's Developmental Age}) \\ & + u_2(\text{Child's Age}) + u_3(\text{Gender}) + e, \end{aligned}$$

where  $\gamma_{00}$  refers to the child-level intercept for an average classroom;  $\gamma_{01}$  represents the maximum likelihood estimate of the population slope estimating average levels of Time 2 COR scores across all classrooms from classroom-level developmental age composition;  $\gamma_{02}$  represents the maximum likelihood estimate of the population slope estimating average levels of Time 2 COR scores across all classrooms from classroom-level size (i.e., the number of students in the classroom) and  $u_0$  represents error in estimating this intercept;  $\gamma_{10}$  represents the maximum likelihood estimate of the population slopes estimating Time 2 COR score from child's developmental age (i.e., Time 1 COR) and  $u_1$  represents error related to that estimate;  $\gamma_{20}$  represents the maximum likelihood estimate of the population slopes estimating Time 2 COR score from child's age and  $u_2$  represents error related to that estimate;  $\gamma_{30}$  represents the maximum likelihood estimate of the population slopes estimating Time 2 COR score from gender and  $u_3$  represents error related to that estimate; and  $\gamma_{11}$  represents the maximum likelihood estimate of the cross-level interaction between child-level developmental age and classroom-level developmental age composition.

Received December 19, 2006

Revision received January 3, 2008

Accepted January 7, 2008 ■



# Peer Victimization and Academic Achievement in a Multiethnic Sample: The Role of Perceived Academic Self-Efficacy

Jochem Thijs and Maykel Verkuyten  
Utrecht University

This study examines the link between perceived peer victimization and academic adjustment in an ethnically diverse sample of 1,895 Grade 6 students nested within 108 school classes. It was hypothesized that students' academic self-efficacy mediates the (negative) link between victimization experiences and academic achievement outcomes. Multilevel analyses were used to test this hypothesis and to explore whether there are differences between ethnic minority and majority group children. Results indicated that peer victimization was negatively associated with both relative class-based, and absolute test-based measures of academic achievement. These associations were similar across different school classes. As expected, the link between victimization and achievement was mediated by perceived academic self-efficacy, suggesting that victimized students did less well academically because they considered themselves to be less competent. The lower perceived self-efficacy of victimized children could be partly attributed to lower global self-esteem and depressed affect. Results were largely similar for ethnic minority and majority group children.

**Keywords:** peer victimization, academic achievement, perceived academic self-efficacy, ethnic minority students

Peer victimization is a considerable problem for a substantial number of children (Juvonen, Nishina, & Graham, 2000). The term *peer victimization* refers to the individual experience of aggressive or negative behaviors by others, including name calling or active exclusion (Hawker & Boulton, 2000; Lopez & DuBois, 2005). There is a bulk of evidence suggesting that victimization experiences can have negative consequences for children's psychosocial functioning. Hawker and Boulton (2000) summarized this evidence in a meta-analysis of cross-sectional studies covering almost 20 years of research. Peer victimization had significant negative effects on all adjustment variables examined in the analysis, including depression, loneliness, global self-esteem, social self-concept, general anxiety, and social anxiety. Although these effects were moderate in size ( $r = .19-.45$ ), the range of affected outcomes attests to the seriousness of the phenomenon (Hawker & Boulton, 2000).

More recently, researchers have focused on the links between peer victimization and academic adjustment, and, in particular, children's achievement outcomes. Studies relying on cross-sectional data have shown that victimized children receive lower grades than their more accepted classmates (Buhs & Ladd, 2001; Graham, Bellmore, & Mize, 2006; Lopez & DuBois, 2005; Schwartz, Farver, Chang, & Lee-Shin, 2002). Further, longitudinal research suggests that these lower achievement outcomes are consequences rather than causes of victimization. It has been shown, for instance, that changes in self-perceived victimization (together

with self-worth and loneliness) are uniquely predictive of GPA (Juvonen et al., 2000), that victimization predicts decreased levels of academic achievement (Buhs, Ladd, & Herald, 2006), and also that academic functioning does not predict changes in peer victimization (Schwartz, Gorman, Nakamoto, & Toblin, 2005). Notwithstanding these research findings, the mechanisms by which victimization affects academic achievement are not fully clear.

Empirical attempts to explain the academic consequences of peer victimization have focused on the role of psychological adjustment. Several studies have shown that maladjustment mediates the negative effects of peer victimization on children's scholastic functioning, including their academic outcomes (Austin & Joseph, 1996; Graham et al., 2006; Juvonen et al., 2000; Lopez & DuBois, 2005; Schwartz et al., 2005). Most of these studies have relied on global or composite measures of emotional well-being, such as depression, loneliness, or anxiety. The findings are consistent with motivational models stating that motivated academic behavior requires a state of emotional well-being (Boekaerts, 1993) or secure relatedness (Connell & Wellborn, 1991; Ryan & Deci, 2000a, 2000b). Clearly, children who are victimized in school do not experience this state, which puts them at risk for unfavorable academic outcomes.

Theoretically, emotional well-being is not the only prerequisite for academic achievement. Both motivational and self-concept theorists posit that self-perceived efficacy is essential as well. Perceptions of efficacy refer to the confidence in one's ability to organize and execute a given course of action or accomplish a task. Proponents of process models of motivation argue that self-directed behaviors and, hence, positive achievement outcomes are dependent on feelings of personal efficacy, in addition to relatedness and autonomy (Connell & Wellborn, 1991; Ryan & Deci, 2000a, 2000b; Skinner, Wellborn, & Connel, 1990). Furthermore,

---

Jochem Thijs and Maykel Verkuyten, Department of Interdisciplinary Social Sciences, Utrecht University, Utrecht, The Netherlands.

Correspondence concerning this article should be addressed to Jochem Thijs, Department of Interdisciplinary Social Sciences, Utrecht University, Heidelberglaan 2, Utrecht 3584 CS, The Netherlands. E-mail: j.t.thijs@uu.nl

it has been concluded that students' academic self-concepts, which include their perceptions of academic efficacy, have reciprocal relations with academic achievement outcomes. Not only are these self-perceptions grounded in actual accomplishments, they also have motivating properties leading to better achievement outcomes (Guay, Larose, & Boivin, 2004; Marsh, Trautwein, Lüdtke, Koller, & Baumert, 2005; Trautwein, Lüdtke, Koller, & Baumert, 2006; Valentine, DuBois, & Cooper, 2004).

There are indications that peer victimization can have a negative impact on children's academic self-efficacy. For instance, negative correlations have been reported between peer victimization experiences and perceived academic competence (Austin & Joseph, 1996; Verkuyten & Thijs, 2002). In addition, Flook, Repetti, and Ullman (2005) found that children's perceived academic self-efficacy was negatively affected by experiences with peer rejection, which is considerably related to peer victimization and has similar correlates (Lopez & DuBois, 2005).<sup>1</sup> A possible reason for these findings is that victimized (or rejected) children receive negative messages about themselves, which negatively affect their overall self-evaluations (Graham & Juvonen, 1998; Hawker & Boulton, 2000), and which may extend to their self-efficacy in the academic domain (Flook et al., 2005).

Given its (anticipated) relations with peer victimization and academic achievement, it is reasonable to expect that academic self-efficacy mediates the link between victimization and achievement. Two of the aforementioned studies provide indirect support for this hypothesis. First, Lopez and DuBois (2005) obtained empirical support for a model in which negative (social and global) self-evaluations mediated the links between peer victimization and peer rejection, on the one hand, and academic as well as behavioral and emotional problems, on the other hand. Although the self-evaluations in that study pertained to the social and global domains rather than to the academic domain, the results indicate that victimization diminishes feelings of efficacy. Second, Flook et al. (2005) examined how peer rejection was related to children's academic performance. It appeared that both perceived academic self-efficacy and internalizing symptoms were independent mediators of this relationship. This indicates that peer rejection and possibly also peer victimization put children at risk for low academic outcomes, not only because it diminishes their sense of emotional well-being but also because it diminishes their sense of academic competence (Flook et al., 2005).

The present research focused on perceived academic self-efficacy as a mediator of the (hypothesized) link between peer victimization and academic achievement. In doing so, we went beyond Flook et al.'s study (2005) in four ways. First, we operationalized students' academic achievement outcomes in two manners. Like Flook et al. (2005), we relied on class-based measures by assessing students' academic accomplishments relative to their classmates. In addition, we included a measure of academic achievement based on standardized tests scores. The use of this score allowed us to examine the impact of victimization beyond the classroom. The distinction between relative and more absolute measures might also be relevant for the mediating role of perceived academic self-efficacy. It has been argued, and found, that academic self-concept has a stronger influence on class-based relative performance than on standardized achievement outcomes (Marsh, 1987; Marsh et al., 2005).

Second, we examined the role of two covariates. Following Flook et al. (2005) who examined the role of internalizing problems, we included a measure of depressed affect to control for diminished levels of well-being associated with victimization. However, we also included a measure of global self-esteem. This allowed us to examine the specific suggestion that the hypothesized link between victimization and perceived academic self-efficacy can be explained as a generalized effect of negative global self-feelings (see Flook et al., 2005; Lopez & DuBois, 2005).

Third, we used a two-level design by examining children nested within a large number of school classes. As a result, we could investigate whether the statistical effects of victimization are similar or different across different classes. Such an examination is of theoretical interest as it can improve our understanding of the potential impact of negative peer treatment. Moreover, it has practical relevance, because it can indicate whether the classroom context should be considered in attempts to prevent or diminish the negative effects of peer victimization on academic outcomes.

The fourth feature of our study was the ethnic diversity of the sample. Like most western countries, the Netherlands hosts a variety of different ethnic groups. It is important that this variety is represented in research. Moreover, use of a multiethnic sample allows examination of whether the links among peer victimization, perceived academic self-efficacy, and achievement outcomes are similar for ethnic majority and ethnic minority group students. Relatively few studies have examined peer victimization and its various effects on children from ethnic minority groups (e.g., Hanish & Guerra, 2000; Storch, Zelman, Sweeney, Danner, & Dove, 2002; Verkuyten & Thijs, 2006). Peer victimization may have different meanings for minority versus majority students. There is evidence, for example, that minority children more often understand victimization experiences as instances of ethnic discrimination (Verkuyten & Thijs, 2000, 2006). It is important to examine these experiences as several studies have found negative effects of ethnic discrimination on children's academic achievement (Graham et al., 2006; Neblett, Philip, Cogburn, & Sellers, 2006; Wong, Eccles, & Sameroff, 2003).

The present study had two goals. First, we examined the relations among peer victimization, perceived academic self-efficacy, and (relative and absolute) academic achievement in an ethnically diverse sample of early adolescents. Our main hypothesis was that perceived academic self-efficacy mediates the (negative) link between children's victimization experiences and their academic achievement outcomes. This hypothesis was tested without and with depressed affect and global self-esteem as covariates. The inclusion of the latter allowed us to evaluate the more specific subhypothesis that peer victimization negatively affects children's academic self-efficacy through negative global self-feelings. The two hypotheses are schematically depicted in Figure 1. The second goal of the study was to explore whether these expected associations hold for both ethnic majority and minority groups.

<sup>1</sup> Whereas peer victimization reflects negative behaviors by individual peers, peer rejection reflects negative attitudes by the peer group (Lopez & DuBois, 2005).



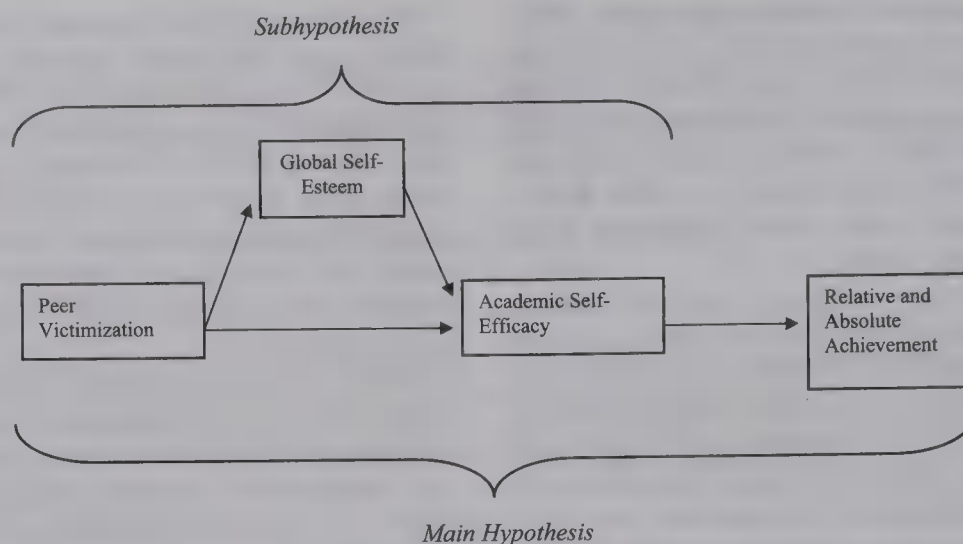


Figure 1. Scheme of hypothesized relations.

## Method

### Participants

Participants were 1,895 Grade 6 students from 108 classes in 81 regular primary schools in the Netherlands. Of these children, 50.6% were girls. According to their ethnic self-definition and the reported ethnicity of their parents, 844 of these children were identified as Dutch. According to the same criteria, 605 children belonged to the three largest minority groups in the Netherlands. They were identified as Turkish ( $n = 299$ ), Moroccan ( $n = 237$ ), or Surinamese ( $n = 69$ ). The remaining 446 children were of mixed or different ethnicities.

### Procedures

All students were tested in the second half of Grade 6 (the spring, i.e., when they already had participated in a standardized achievement test; see next section). Each of them completed a questionnaire under supervision and simultaneously within their classrooms. In the introduction to this questionnaire, students were asked to answer questions about school and themselves, and their anonymity was guaranteed. Almost all students were able to complete the questions within 40 min. The measures of interest to the present study were included in the following fixed order: peer victimization, global self-esteem, academic self-efficacy, depressed affect, relative achievement, and absolute achievement.

There were no missing variables for 95% of the cases. For the remaining children, 2%–9% of the values were missing ( $Mdn = 2\%$ ). We imputed these scores using the expectation maximization algorithm. This procedure is adequate when values are missing at random (Bernaards & Sijtsma, 1999).

### Measures

**Peer victimization.** Perceptions of peer victimization were assessed with four items, which referred to the frequencies of being teased or called names and the frequencies of being excluded in the school and neighborhood. These items were developed by the authors from Dutch research on early adolescents' own understanding of peer victimization (Verkuyten, Kinket, & van der

Wielen, 1997). In previous studies, these items were found to be negatively related to children's global self-esteem (Verkuyten & Thijs, 2001) and depressed affect (Verkuyten, 2003), supporting their concurrent validity. The items were scored on a scale ranging from 1 (*no, never*) to 5 (*yes, very often*). Principal components analysis (PCA) on the items revealed one component that explained 54% of the variance. Cronbach's alpha (for internal consistency) was .72.

**Academic self-efficacy.** Perceived academic self-efficacy was assessed with four items adapted from the scholastic competence scale of Harter's (1988) Self-Perception Profile for Adolescents (SPPA). The SPPA is an established self-concept measure containing eight subscales for domain-specific self-evaluations for which adequate internal consistencies (Cronbach's alpha's  $> .73$ ) and strong factorial validity have been reported (Harter, 1988). Harter (1988) did not report external validity findings for the SPPA (see also Byrne, 1996). However, studies using adaptations of this measure have reported moderate relations between the scholastic competence scale and students' academic grades ( $r > .34$ ; Soenens & Vansteenkiste, 2005; Wichstrøm, 1995).

In the SPPA, respondents are required to choose one statement from a pair of two opposite statements and then to express their level of agreement with their choice. However in the present research, as in other studies (Soenens & Vansteenkiste, 2005; Wichstrøm, 1995), children were presented with single statements rather than paired statements. We chose this format because in previous studies, we found that many children had difficulties with the paired format. Agreement with the items was rated on a scale ranging from 1 (*no, certainly not*) to 4 (*yes, certainly*). For example, children were asked whether they often forget what they learn and whether they are able to learn very well. Cronbach's alpha was .62 for this scale. The items loaded on one component that explained 47% of the variance.

**Relative achievement.** Elementary school children in the Netherlands receive their grades from their teachers. As in other countries, these grades are based (in part) on students' achievements relative to those of their classmates. We collected information on perceived relative academic position within the classroom by means of three Willig Scales (see Burns, 1979). The Willig Scale

is a self-anchoring, 11-step rating scale that has been used in previous studies among ethnic majority and minority early adolescents in the Netherlands (Verkuyten, Thijs, & Canatan, 2001). The top of the scale (10) marks the best performing student in one's class (i.e., the student receiving the highest grades in class), and the lowest step (0) marks the worst performing student. Children were asked to use this scale to rate their general performance, their achievement in language learning, and their achievement in mathematics. These three ratings loaded on one component accounting for 69% of the variance, and they yielded a Cronbach's alpha of .75. Students' mean score on this relative achievement measure was 6.72 ( $SD = 1.60$ ). Its distribution was slightly skewed to the left ( $-.424, p < .001$ ) but it had no significant kurtosis. Supporting its concurrent validity, previous studies found that this measure was positively related to academic motivation, academic competence, and school satisfaction (Verkuyten & Thijs, 2002; Verkuyten et al., 2001).

**Absolute achievement.** To obtain a more absolute measure of their academic achievement, we used students' self-reports of their official secondary school advice. In the Netherlands, students receive their secondary education advice from their teachers in the final grade (Grade 6) of primary school. Teachers take several considerations into account when giving this advice. However, the advice is predominantly based on students' scores on a standard national school achievement test (CITO test) and is highly correlated with these scores ( $r > .85$ ; Driessen & Doesborgh, 2005; Kapinga, 2002). Thus, the educational advice is a valid measure of students' academic achievement.

The secondary Dutch education system has five levels: a) initial professional education, b) general and vocational education, c) senior general secondary education, d) university preparatory education atheneum, and e) university preparatory education gymnasium (high-level grammar school). Teachers' advice to children involves one type of education or the combination of two bordering levels of education. Our absolute achievement measure was a 7-point scale including each level and the combinations of b and c, and c and d.<sup>2</sup> The distribution of this scale had no significant kurtosis. However, it was slightly skewed to the right (.918,  $p < .001$ ). The mean score on this measure was 2.73 ( $SD = 1.74$ ), and the median was 2.<sup>3</sup>

**Depressed affect.** Three items that inquired about sadness, nervousness, and fear were used to measure students' depressed affect. Cronbach's alpha was .62. The items were taken from the Profile of Mood States (McNair, Lorr, & Droppleman, 1971) on the bases of their face validity. We used the same 4-point response format (from 1, *no, certainly not* to 4, *yes, certainly*) used for the Academic Self-Efficacy Scale. PCA on these three items yielded one component that explained 58% of the variance.

**Global self-esteem.** Global self-esteem was assessed with items from the well-known 10-item Rosenberg (1965) Self-Esteem Scale. Early adolescents have been found to have difficulties in responding to negatively worded self-esteem items (Marsh, 1986). Therefore, we used the five positively worded items of the Rosenberg scale. In addition, each item had the same four-point response format as the perceived academic self-efficacy measure. Cronbach's alpha for these five items was .75. The items loaded on one component that explained 52% of the variance. In support of the concurrent validity of the scale, previous Dutch research has shown that this abbreviated measure is negatively related to de-

pressed affect (Verkuyten, 2003) and positively to self-concept stability (Verkuyten, 1995) and ethnic self-regard (Verkuyten & Thijs, 2004).

### Data Analysis

Participants were sampled through their classes rather than individually. As children attending the same class tend to be alike in some respects, data for individual participants were probably not independent. When dependent data are analyzed with conventional statistical tests, standard errors are underestimated, and results may be spuriously significant (Snijders & Bosker, 1999). This can be prevented with multilevel analyses. Multilevel analysis can correct for dependencies between observations for individual subjects (e.g., pupils) nested within the same units (e.g., classes). Moreover, it can be used to analyze variable numbers of subjects per unit (Snijders & Bosker, 1999). In this study, we conducted multilevel analyses with MLwiN Version 2.0 (Rasbash, Browne, Healy, Cameron, & Charlton, 2004) using the iterative generalized least squares algorithm. Two levels were specified: Level 1 pertaining to individual differences within classes and Level 2 pertaining to differences between classes.<sup>4</sup>

The measures of relative and absolute achievement were strongly related (see Table 1). To examine whether both measures were similarly affected by the independent variables, we analyzed them simultaneously in multivariate multilevel models. For this purpose, both measures were standardized, and an additional level was specified. This level, Level 0, was included to define the multivariate structure (Goldstein, 1995; Snijders & Bosker, 1999). All other variables were examined with univariate multilevel models.

The multivariate multilevel model is an extension of the univariate model. A univariate two-level regression model with one fixed Level 1 predictor  $x$  can be expressed by  $y_{ij} = \beta_0 + \beta_1 x_{ij} + e_{ij} + u_j$ , with  $\text{var}(e_{ij}) = \sigma_e^2$  and  $\text{var}(u_j) = \sigma_u^2$ . In this equation, the subscripts  $i$  and  $j$  denote units at Level 1 (e.g., students) and Level 2 (e.g., classes), respectively;  $\beta_0$  is the intercept and  $\beta_1$  is the slope; and  $e_{ij}$  and  $u_j$  are the residuals for each level. The two-variate variant of the univariate model is represented by  $y_{hij} = \beta_{01}z_{1hij} + \beta_{02}z_{2hij} + \beta_{11}z_{1hij}x_i + \beta_{12}z_{2hij}x_i + e_{1ij}z_{1hij} + e_{2ij}z_{2hij} + u_{1j}z_{1hij} + u_{2j}z_{2hij}$ , with  $\text{var}(e_{1ij}) = \sigma_{e1}^2$ ,  $\text{var}(e_{2ij}) = \sigma_{e2}^2$ ,  $\text{cov}(e_{1ij}, e_{2ij}) = \sigma_{e1e2}$ ,  $\text{var}(u_{1j}) = \sigma_{u1}^2$ ,  $\text{var}(u_{2j}) = \sigma_{u2}^2$ , and  $\text{cov}(u_{1j}, u_{2j}) = \sigma_{u1u2}$ . Here, the additional level (Level 0) is indicated by the subscript  $h$ . In

<sup>2</sup> Within this scaling, the lowest two levels and the highest two levels are relatively close to each other, and many children are given advice that combines Levels 2 and 3, or Levels 3 and 4. For these reasons, it was decided to include these combinations as separate scale-points.

<sup>3</sup> Because both achievement measures were skewed, we looked for outliers. There were no extreme scores on the absolute measure. However, for five cases, relative achievement scores were more than 3.29 standard deviations (maximum  $SD$ , 4.21) below the sample mean corresponding to  $p < .001$  (Tabachnick & Fidell, 1996). These cases were retained in the analyses because analyses without them yielded virtually the same results, and sample size was large.

<sup>4</sup> The data had a three-level structure with students nested in classes nested in schools. However, three-level models could not be examined because class and school levels were heavily confounded as 60 schools (74%) were represented by only one class each. Most Dutch primary schools have only one class for each grade.



Table 1  
*Intercorrelations, Means, and Variance Components of All Variables*

Variable	1	2	3	4	5	6	M	Variance		
								Level 1	Level 2	Total
1. Peer victimization	—						1.98	.466	.019**	.485
2. Academic self-efficacy	-.25**	—					2.98	.263	.007*	.270
3. Relative achievement	-.16**	.53**	—				0.00	.977	.023*	1.00
4. Absolute achievement	-.10**	.43**	.58**	—			0.00	.912	.090**	1.00
5. Depressed affect	.36**	-.35**	-.17**	-.09**	—		2.05	.315	.013**	.328
6. Global self-esteem	-.22**	.32**	.25**	.07**	-.18**	—	3.07	.292	.022**	.314

Note. Both relative and absolute achievement measures were standardized.

\*  $p < .01$ . \*\*  $p < .001$ .

addition,  $z_{1ij} = 1$  and 0 for the first and the second dependent variables, respectively, and  $z_{2ij} = 1 - z_{1ij}$  (see Goldstein, 1995).

For both achievement measures, regression analyses proceeded in two steps. First, the statistical effect of each independent variable was constrained to be similar for both relative and absolute achievement (e.g.,  $\beta_{11}z_{1hij}x_i = \beta_{12}z_{2hij}x_i$ ). Second, if it significantly improved the fit of the model ( $p < .05$ ), this constraint was released, and different effects were estimated for both achievement variables. Unless otherwise indicated, the effects on relative and absolute achievement were different in all analyses.

## Results

### Preliminary Analyses

Before examining the (unique) statistical effects of peer victimization and the mediating role of perceived academic self-efficacy, we estimated the intercorrelations and the variance distributions of all variables. Correlations are shown in the left part of Table 1. As expected, peer victimization was positively related to depressed affect and negatively related to all other variables. These correlations ranged from small to moderate (Cohen, 1988). The relations between the remaining variables were moderate to large, except for the correlations of depressed affect with both achievement variables and global self-esteem and the correlation of global self-esteem with absolute achievement. The right part of Table 1 contains the means and variance components of all variables. The latter were obtained by means of the so-called intercept-only regression model. This model provides an estimate of the intraclass correlation coefficient ( $\rho$ ), which represents the proportion of variance at Level 2 (the class level) for each dependent variable (Snijders & Bosker, 1999). All variables had significant variance at Level 2, indicating systematic differences between classes on these measures. For peer victimization, academic self-efficacy, depressed affect, and relative achievement, these differences constituted a relatively small portion of their total variance: 3.9%, 2.6%, 4.0%, and 2.3%, respectively. For global self-esteem and absolute achievement, the proportion of Level 2 variance was considerably higher. It appeared that 7.0% and 9.0%, respectively, of the individual differences on these variables could be attributed to differences between classes.

### Statistical Effects of Victimization on Academic Functioning

The correlations in Table 1 indicate significant links between peer victimization, on the one hand, and perceived academic self-efficacy and relative and absolute achievement, on the other. However, the two-level structure of our data is not reflected in these correlations. To properly examine whether students' victimization experiences affect within- and between-class differences in the academic variables, we performed multilevel regression models for self-efficacy and both achievement variables. In these models, victimization was included as a single predictor. Its effects were estimated in two steps. First, the effect of victimization was fixed across all classes (Level 2). Second, the regression slopes for victimization were allowed to vary randomly across Level 2. The second step did not result in significant model improvement ( $p > .10$ ) indicating similar effects across all classes. Hence, only the fixed effects of victimization were inspected. These effects are shown in Table 2 (Models 1 and 2).<sup>5</sup>

Consistent with the correlations in Table 1, the analyses revealed that victimization had negative effects on all three academic variables, explaining 6.3% of the variance in perceived academic self-efficacy and 2.7% and 0.8% of the variance in relative and absolute achievement, respectively. Table 2 also shows deviance statistics, which can be used to compare the fit of nested models. Differences between these statistics follow a chi-square distribution, and degrees of freedom are given by the differences in numbers of parameters (Snijders & Bosker, 1999). As shown in Table 2, Models 1 and 2 were significant improvements on the intercept-only models for academic self-efficacy and for relative and absolute achievement.

Given the negative effects of peer victimization and the significant correlations between perceived academic self-efficacy and both relative and absolute achievement (Table 1), it was appropri-

<sup>5</sup> Gender was not included in any of the reported multilevel analyses. Additional analyses, not reported here, indicated significant gender differences for academic self-efficacy, relative achievement, global self-esteem, and depressed affect, with higher means for boys on the first three measures ( $p < .01$ ) and lower scores on depressed affect. However, gender did not interact with victimization, self-efficacy, self-esteem, or depressed affect in the prediction of achievement, nor did its inclusion substantially alter the statistical effects of these predictors.

Table 2  
*Multilevel Effects of Peer Victimization on Academic Self-Efficacy, Relative Achievement, and Absolute Achievement*

Predictor	Model 1	Model 2		Model 3	
	Academic self-efficacy	Achievement		Achievement	
		Relative	Absolute	Relative	Absolute
Constant	3.347	0.499	0.249	-2.831	-2.563
Peer victimization	-0.186**	-0.243**	-0.138**	-0.058 <sup>†</sup>	0.018
Academic self-efficacy	—	—	—	0.995**	0.841**
Variance					
Level 1	.247	.945	.900	.699	.725
Level 2	.006	.028	.091	.024	.090
Total (% variance explained)	.253 (6.3)	.973 (2.7)	.992 (0.8)	.723 (27.7)	.815 (18.5)
Deviance	2763.562	9693.010		9060.759	
$\chi^2$	121.893** <sup>a</sup>	55.121** <sup>b</sup>		632.251** <sup>b</sup>	
Reference model	Intercept—only	Intercept—only		2	

<sup>a</sup> Degree of freedom = 1. <sup>b</sup> Degrees of freedom = 2.

\*\*  $p < .001$ . <sup>†</sup>  $p < .05$ .

ate to examine whether self-efficacy mediated the statistical effects of victimization on the achievement measures (see Baron & Kenny, 1986).

The critical test for mediation is that the influence of the predictor on the dependent variable is substantially reduced when the mediator is added as an additional predictor. To conduct this test and to evaluate our main hypothesis (see Figure 1), we regressed both achievement measures on victimization and academic self-efficacy. The result is displayed in the right part of Table 2 (Model 3). Perceived academic self-efficacy had a positive effect on both relative and absolute achievement. This effect was stronger for the first than for the latter measure ( $p < .01$ ) but accounted for more than 17.8% of unique variance in both cases. When the influence of perceived self-efficacy was partialled out, the effects of victimization were considerably reduced. Further

analyses revealed that victimization could only explain an additional 0.2% of the variance in relative achievement compared with 2.7% when perceived self-efficacy was not included (Table 2). In addition, the effect on absolute achievement was no longer significant, implying complete mediation.

#### *Analyses with Covariates*

To investigate whether the aforementioned results were upheld independent of depressed affect and global self-esteem, we added these covariates as predictors to the regression equations in Table 2. First, academic self-efficacy and the achievement variables were regressed on both peer victimization and the covariates. Results are shown under Models 4 and 5 in Table 3. Depressed affect had negative unique effects on self-efficacy and relative achievement,

Table 3  
*Multilevel Effects of Peer Victimization Controlled for Covariates*

Predictor	Model 4	Model 5		Model 6	
	Academic self-efficacy	Achievement		Achievement	
		Relative	Absolute	Relative	Absolute
Constant	2.926	-0.507	0.014	-3.262	-2.739
Peer victimization	-0.073**	-0.120**	-0.096*	-0.052	-0.030
Depressed affect (Covariate 1)	-0.247**	-0.193**	-0.075	0.041	0.158**
Global self-esteem (Covariate 2)	0.229**	0.375**	0.100 <sup>†</sup>	0.159**	-0.117*
Academic self-efficacy	—	—	—	0.943**	0.943**
Variance					
Level 1	.211	.893	.897	.695	.715
Level 2	.004	.022	.091	.020	.091
Total (% variance explained)	.215 (20.4)	.915 (8.5)	.988 (1.2)	.715 (28.5)	.806 (19.4)
Deviance	2466.951	9558.870		8993.166	
$\chi^2$	296.611** <sup>a</sup>	134.140** <sup>b</sup>		565.704** <sup>a</sup>	
Reference model	1	2		3	

<sup>a</sup> Degrees of freedom = 2. <sup>b</sup> Degrees of freedom = 4.

\*  $p < .01$ . \*\*  $p < .001$ . <sup>†</sup>  $p < .05$ .



and global self-esteem had positive unique effects on all variables. More important, the effects of victimization were still significant, but they appeared to be smaller compared with those of Models 1 and 2 (Table 2). This seemed to support our subhypothesis that the negative link between peer victimization and self-efficacy could be partly attributed to negative overall self-feelings (see Figure 1).

The Sobel test for mediation was used to examine the indirect effects of peer victimization. An indirect effect of a variable  $x$  on a variable  $y$  through a variable  $z$  can be expressed as  $ab$ , with  $a$  being the effect of  $x$  on  $z$  and  $b$  being the effect of  $z$  on  $y$ . The Sobel test provides a  $z$  statistic for this indirect effect by dividing it by its estimated standard error  $(b^2 s_a^2 + a^2 s_b^2)^{1/2}$ ; see Baron & Kenny, 1986; MacKinnon, Warsi, & Dwyer, 1995). The Sobel test revealed significant indirect effects through global self-esteem and also depressed affect (respectively,  $z = 7.34$ , and  $z = -9.89$ ,  $p < .001$ ). Hence, these variables accounted for part of the relationship between peer victimization and academic self-efficacy.

Next, relative and absolute achievement were regressed on peer victimization, depressed affect and global self-esteem, and academic self-efficacy. Results are displayed under Model 6 (Table 3). Depressed affect had no effect on relative achievement but had a positive (rather than a negative) effect on absolute achievement. Global self-esteem was positively related to relative achievement but had a negative effect on absolute achievement. More important, however, perceived self-efficacy was still a strong significant predictor of both achievement measures (with similar effects on relative and absolute achievement). Thus, it appeared that our previous mediation findings (in Model 3, Table 2) could not be reduced to effects of depressed affect or global self-esteem.

### Ethnic Differences

We examined whether these results applied to ethnic minority (Turkish-Dutch, Moroccan-Dutch, and Surinamese-Dutch) and majority (Dutch) students. In doing so, three dummy variables were created: TUR, MOR, and SUR. These variables were coded, respectively, as 1 for the Turkish-Dutch and 0 for the other children, 1 for the Moroccan-Dutch and 0 for the other children, and 1 for the Surinamese-Dutch and 0 for the other children. When included together, these dummies represented the difference between each of the three minority groups and the group of Dutch children.

Prior to examining the statistical effects of victimization for the different ethnic groups, we examined whether minority and majority students reported similar levels of peer victimization and academic self-efficacy. To this aim, we regressed these variables on the three dummy variables. For peer victimization, 1.5 % of the variance could be attributed to students' ethnic group. Compared with the Dutch students, Turkish and Moroccan students reported fewer instances of peer victimization (respectively,  $b = -.104$ ,  $p < .05$ , and  $b = -.267$ ,  $p < .001$ ). The difference between Surinamese and Dutch students was not significant. For perceived academic self-efficacy, there were no significant differences between the Dutch and the other students.

Next, we examined whether the links between victimization, and perceived academic self-efficacy and both achievement measures differed for minority versus majority students. These variables were regressed on the three dummy variables, on peer victimization, and on the three interactions between the former and

the latter. Inspection of the resulting models revealed that none of the interactions was significant, which indicates that victimization works similarly for the Turkish, Moroccan, and Surinamese as compared with the Dutch students:  $b = -.178$ ,  $p < .001$ , for academic self-efficacy, and  $b = -.186$  (similar) for both achievement variables,  $p < .001$ .

Subsequently, we examined whether for both the minority and the majority students, perceived self-efficacy mediated the relationship between victimization and relative and absolute achievement. Both achievement measures were regressed on the three dummy variables, victimization, academic self-efficacy, and the interactions between the dummy variables and academic self-efficacy. The result is shown under Model 7 in Table 4. Perceived academic self-efficacy and its interactions with the dummy variables were significant predictors of academic achievement. Their effects differed for the relative versus the absolute measure ( $p < .01$ ). Self-efficacy had stronger effects on the relative achievement of the Dutch versus the Turkish and Moroccan students, and the absolute achievement of the Dutch versus all minority children. However, further inspection of the data revealed that the effects of self-efficacy were positive for all ethnic groups: for relative achievement,  $b_{\text{Turks}} = .776$ ,  $b_{\text{Moroccans}} = .695$ ,  $b_{\text{Surinamese}} = .834$ , and  $b_{\text{Dutch}} = 1.200$ ,  $p < .001$ , and for absolute achievement,  $b_{\text{Turks}} = .509$ ,  $b_{\text{Moroccans}} = .709$ ,  $b_{\text{Surinamese}} = .479$ , and  $b_{\text{Dutch}} = .995$ ,  $p < .05$ .<sup>6</sup> When these effects of perceived self-efficacy were partialled out, the effect of victimization was no longer significant. Thus, although academic self-efficacy had different effects of minority versus majority children, it mediated the effects of victimization for all of them.

Finally, we examined whether the effects of victimization as well as the mediation findings were upheld independent of depressed affect and global self-esteem. First, we regressed academic self-efficacy and the achievement variables on the covariates, the dummy variables, and peer victimization. As in the total sample, the effects of peer victimization were significant but also weaker:  $b = -.070$ ,  $p < .001$ , for perceived academic self-efficacy, and  $b = -.186$ ,  $p < .01$ , for both achievement variables. Next, we added depressed affect and global self-esteem to the mediation model (Table 4). As shown under Model 8, the results for self-efficacy and its interactions with the dummy variables were unaffected by the inclusion of these covariates.

### Discussion

This study examined the associations between peer victimization and academic achievement in a large sample of early adolescents. The research had a cross-sectional design, and, hence, our findings do not allow causal conclusions. However, our analysis and interpretation of the direction of effects is consistent with theoretical expectations and with longitudinal findings of victimization being a cause rather than a consequence of low achievement outcomes (Buhs et al., 2006; Juvonen et al., 2000; Schwartz et al., 2005).

<sup>6</sup> The effects for the Dutch children can be directly inferred from Models 7 and 8. The statistical effects for the minority children can be obtained by adding the regression coefficient for each minority group to the coefficient for the Dutch children.

Table 4  
*Mediation Analyses for Relative and Absolute Achievement Among Minority and Majority Students*

Predictor	Model 7		Model 8	
	Achievement		Achievement	
	Relative	Absolute	Relative	Absolute
Constant	-3.524	-2.768	-4.050	-2.837
TUR	1.325**	1.020*	1.235**	0.947*
MOR	1.571**	0.536	1.414**	0.503
SUR	1.034	1.277	0.970	1.272
Peer victimization	-0.031	-0.031	-0.050	-0.050
Academic self-efficacy	1.200**	0.995**	1.174**	1.040**
Academic self-efficacy—TUR	-0.424**	-0.486**	-0.409**	-0.463**
Academic self-efficacy—MOR	-0.505**	-0.286 <sup>†</sup>	-0.468**	-0.269 <sup>†</sup>
Academic self-efficacy—SUR	-0.366	-0.516 <sup>†</sup>	-0.351	-0.512 <sup>†</sup>
Depressed affect (Covariate 1)	—	—	0.102*	0.102*
Global self-esteem (Covariate 2)	—	—	0.146**	-0.077
Variance				
Level 1	.689	.682	.684	.676
Level 2	.025	.077	.025	.078
Total (% variance explained)	.714 (28.6)	.759 (24.1)	.709 (29.1)	.754 (24.6)
Deviance	6808.043		6773.286	

Note. TUR is a dummy variable, with Turkish = 1 and Other = 0; MOR is a dummy variable, with Moroccan = 1 and Other = 0; SUR is a dummy variable, with Surinamese = 1 and Other = 0; For both peer victimization and depressed affect, common coefficients were estimated.

\*  $p < .01$ . \*\*  $p < .001$ . <sup>†</sup>  $p < .05$ .

As expected, students who reported more victimization experiences had less favorable achievement outcomes. This finding is in agreement with the results of other studies showing negative links between peer victimization and academic adjustment. These other studies have either relied on relative measures of achievement such as GPAs or teacher ratings (Graham et al., 2006; Juvonen et al., 2000; Lopez & DuBois, 2005; Schwartz et al., 2002) or on standardized achievement measures (Buhs & Ladd, 2001; Buhs et al., 2006). In the present study, students' relative and absolute academic achievements were simultaneously analyzed. Hence, we could examine the impact of victimization both within and beyond the classroom. Our results indicate that victimization had a stronger statistical effect on the class-based relative achievement measure as compared with the absolute measure. Thus, the impact of victimization was most pronounced for children's accomplishments relative to their classmates. Still, this influence was not confined to students' relative achievement but extended to their official secondary education advice, which can be considered a strong indicator of their absolute achievement (Driessen & Doesborgh, 2005; Kapinga, 2002). Although the links between victimization and achievement are not very strong, they are important because of the potential influence of victimization experiences on students' future academic and occupational careers.

The multilevel structure of the data allowed us to examine peer victimization and its impact within and between classes. As in previous research (Verkuyten & Thijs, 2000), there were systematic between-class differences in the level of peer victimization. However, the effects of victimization were similar across the different classes. This suggests that victimization experiences have similar meanings for children inhabiting different classrooms and

also that shared classroom factors do not affect these meanings. Of course, this is not to say that victimization cannot have different consequences for individual students. Rather, the findings indicate that any practical attempt to prevent lower achievement as a consequence of peer victimization should focus on the characteristics and needs of individual students.

In support of our main hypothesis, the negative associations between perceived victimization and achievement were mediated by perceptions of lower academic self-efficacy. In agreement with previous findings (Flook et al., 2005; Verkuyten & Thijs, 2002), this suggests that children who experience higher rates of peer victimization consider themselves to be less academically competent. This link was independent of global self-esteem and depressed affect. Consistent with our subhypothesis, self-esteem explained a significant part of the link between victimization and self-efficacy. However, this link still existed when its influence was partialled out, and thus, the low perceived academic self-efficacy among victimized students could only be partly attributed to general negative self-evaluations (cf., Flook et al., 2005; Lopez & DuBois, 2005). This raises the question of the specific process behind the association between peer victimization and academic self-efficacy. Perhaps it is not so much global self-esteem but, rather, general self-efficacy that mediates this link. General self-efficacy and global self-esteem are strongly related constructs but the latter emphasizes affective aspects of the self and the former refers to confidence in one's ability to accomplish tasks (Chen, Gully, & Eden, 2004). Another explanation is that the negative messages about themselves that children receive in peer victimizations may involve their intellectual and academic abilities (e.g.,



“stupid,” “dumb”). Future studies are needed to explore these interpretations.

Students' perceived self-efficacy was related to their academic outcomes. This finding supports the notion that perceived self-efficacy has motivating properties leading to better achievement outcomes (Connell & Wellborn, 1991; Marsh et al., 2005; Ryan & Deci, 2000a, 2000b; Skinner et al., 1990; Trautwein et al., 2006). Moreover, our results are consistent with Marsh et al.'s (2005) conclusion that perceived academic self-concept has a stronger influence on school-based performance measures than on standardized achievement outcomes.

Our finding that depressed affect did not affect the mediating role of perceived self-efficacy is consistent with the work of Flook et al. (2005) who showed that self-efficacy was a unique predictor of school outcomes independent of children's internalizing symptoms. However, contradictory to the results of these researchers, depressed affect had no unique negative statistical effects on academic achievement in the present study. Perhaps, this was due to measurement differences. Flook et al. (2005) used a broad, 26-item measure for internalizing symptoms, which did not only entail depressed affect and anxiety but also withdrawal and somatic complaints. Moreover, they relied on teacher reports, rather than self-reports, to assess these symptoms. The absence of a unique, negative link between depressed affect and achievement might seem inconsistent with models of motivation that hold that self-directed behaviors are dependent on emotional well-being (Boekaerts, 1993) or on feelings of self-efficacy, autonomy, and relatedness (Connell & Wellborn, 1991; Ryan & Deci, 2000a, 2000b; Skinner et al., 1990).<sup>7</sup> Yet, it should be noted that our study was not intended to test these models and that our measure of depressed affect was probably too narrow to represent the concepts of well-being or relatedness. On the basis of the present findings, it seems reasonable to conclude that academic self-efficacy predicts students' achievement independent of their socioemotional welfare. Still, future studies are needed to further support this conclusion.

An important feature of our study was the ethnic diversity of the sample. This allowed us to examine the impact of victimization on the academic adjustment of ethnic minority (Turkish-Dutch, Moroccan-Dutch, and Surinamese-Dutch) as compared with majority (Dutch) students. Our results show that the role of victimization and the mediation by academic self-efficacy were independent of minority status. Therefore, our findings support a “one model fits all” approach to studying the academic adjustment of early adolescents in multiethnic settings. However, although there were no ethnic differences in the links among victimization, perceived self-efficacy, and academic achievement, the associations between the latter two constructs were weaker in the minority than the majority samples. We do not have a clear-cut explanation for this finding. It is possible, however, that the weaker associations among the minority students reflect the process of psychological disidentification, which has been found among negatively stereotyped minority groups (Major, Spencer, Schmader, Wolfe, & Crocker, 1998; Steele, 1997). This process involves the disengagement of self-evaluations from academic accomplishments in order to protect one's self-worth. Following the bidirectional link between academic achievement and perceived self-efficacy (see Marsh et al., 2005; Trautwein et al., 2006), one could argue that disidentification works both ways. That is to say, once minority

students detach their perceived self-efficacy from their academic outcomes, these perceptions will have less motivating properties (Marsh et al., 2005; Trautwein et al. 2006; Valentine et al., 2004). Longitudinal studies are needed to test this idea. Irrespective of the exact explanation, the findings further emphasize the usefulness of multiethnic samples in studies of the school adjustment in early adolescent students (see Hanish & Guerra, 2000; Storch et al., 2002).

To evaluate the present research, the reader should consider several qualifications. First, as noted, our design was cross-sectional, and, thus, the possibility of inverse or reciprocal effects cannot be ruled out. Future studies should use longitudinal designs to examine our hypotheses. Still, there are arguments in favor of our interpretation. As mentioned, there are longitudinal findings showing that victimization is more a cause than a consequence of academic achievements (Buhs et al., 2006; Schwartz et al., 2005). Furthermore, there are longitudinal findings showing that academic self-concept influences educational attainment level (Guay et al., 2004). Finally, a set of additional analyses, not reported here, indicated that the link between victimization and perceived self-efficacy was not mediated by students' achievement outcomes. This suggests that if victimization influenced students' academic adjustment, as argued in the present study, it affected students' perceived self-efficacy prior to their actual outcomes.

Second, the study was limited by its reliance on student reports. Whereas the self-report method was adequate for the assessment of academic self-efficacy, self-esteem, and depressed affect, it is possible that the achievement and victimization measures were affected by response bias. Future studies should obtain achievement data from teachers or school records and could assess victimization through sociometric ratings or teacher reports. However, we agree with other researchers that perceptions of victimization should be studied because of phenomenological reasons and their psychological consequences (Graham & Juvonen, 1998). In addition, there is reason to assume that students' reports of their academic achievement were valid and, hence, reliable. Previous research has found that students' self-reported grades were strongly related to their actual GPA (Dornbusch, Ritter, Leiderman, Roberts, & Fraleigh, 1987).

Third, absolute achievement was measured with the students' official secondary educational advice. Research has found that this advice is highly correlated ( $r > .85$ ) with scores on the national standard school achievement test (Driessen & Doesborgh, 2005; Kapinga, 2002). However, the educational advice does not always correspond fully to the score on this test. Differences between educational advice and the test score could be due to the fact that teachers take noncognitive factors into consideration, such as the pupil's motivation and the wishes of the parents and the child (Driessen & Doesborgh, 2005). However, 70% of the variation in the educational advice can be attributed to students' language, math, and reading achievements, and ethnic differences in the educational advice disappear when the effects of these achieve-

<sup>7</sup> Unexpectedly, depressed affect had positive statistical effects on absolute achievement in Tables 3 and 4, and global self-esteem had a negative statistical effect on absolute achievement in Table 3. We do not have an explanation for these findings. However, these statistical effects were small, explaining 0.4%, 0.0%, and 0.5%, respectively, of unique variance.



ment scores are controlled (Driessen & Doesborgh, 2005). Hence, the secondary educational advice appears to be a valid measure of academic achievement.

Fourth, our operationalization of depressed affect was limited. This variable was assessed with three items only yielding a moderate degree of internal consistency. As noted, the unique, negative effects of this measure on students' academic achievement should be interpreted with care. Yet, we think that by including depressed affect in our design, we were able to draw firmer conclusions about the unique mediating role of perceived academic self-efficacy.

Finally, the present study did not consider other potentially important factors that bear upon students' academic achievement. For instance, controlling for students' actual cognitive abilities would have strengthened our conclusions about the mediating role of academic self-efficacy. Future research should examine how victimization affects students' academic adjustment next to, or in interaction with, characteristics such as their abilities, aspects of their home environments, their schools' (instructional) climates, and their relationships with teachers. Still, our finding that victimization had similar statistical effects across different classes and different ethnic groups suggests that its influence is rather uniform.

Despite these limitations, we think that the present study makes a contribution to the literature by examining whether peer-victimized students do less well academically due to self-perceptions of academic incompetence. The findings support the mediating role of perceived self-efficacy. It was found that this role cannot be attributed to general negative self-feelings and is similar for both ethnic minority and majority groups. Experiences of peer victimization appear to have various negative effects for children, including lower academic achievement. Children who have to deal with peer victimization tend to feel academically less competent and thereby miss an important motivation to perform and achieve.

## References

- Austin, S., & Joseph, S. (1996). Assessment of bully/victim problems in 8 to 11 year-olds. *British Journal of Educational Psychology*, 66, 447–456.
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182.
- Bernaards, C. A., & Sijtsma, K. (1999). Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse. *Multivariate Behavioral Research*, 34, 277–313.
- Boekaerts, M. (1993). Being concerned with well-being and with learning. *Educational Psychologist*, 28, 149–167.
- Buhs, E., & Ladd, G. (2001). Peer rejection as antecedent of young children's school adjustment: An examination of mediating processes. *Developmental Psychology*, 37, 550–560.
- Buhs, E. S., Ladd, G. W., & Herald, S. L. (2006). Peer exclusion and victimization: Processes that mediate the relation between peer group rejection and children's classroom engagement and achievement? *Journal of Educational Psychology*, 1, 1–13.
- Burns, R. (1979). *The self-concept: Theory, measurement, development, and behaviour*. London: Longman.
- Byrne, B. M. (1996). *Measuring self-concept across the life span*. Washington, DC: American Psychological Association.
- Chen, G., Gully, S. M., & Eden, D. (2004). General self-efficacy and self-esteem: Toward theoretical and empirical distinction between correlated self-evaluations. *Journal of Organizational Behavior*, 25, 375–395.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.
- Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy, and relatedness: A motivational analysis of self-system processes. In M. R. Gunnar & L. A. Sroufe (Eds.), *Self processes and development* (Vol. 23, pp. 43–77). Hillsdale, NJ: Erlbaum.
- Dornbusch, S. M., Ritter, P. L., Leiderman, P. H., Roberts, D. F., & Fraleigh, M. J. (1987). The relation of parenting style to adolescent school performance. *Child Development*, 58, 1244–1257.
- Driessen, G., & Doesborgh, J. (2005). Relaties tussen achtergrondkenmerken en competenties van leerlingen en hun advies voor voortgezet onderwijs. [Relations between students' background characteristics and their secondary educational advice]. In G. Driessen, J. Doesborgh, G. Ledoux, M. Overmaat, J. Roeleveld, & I. van der Veen (Eds.), *Van basis-naar voortgezet onderwijs. Voorbereiding, advisering en effecten* [From primary to secondary education: Preparation, advice, and impact] (pp. 39–79). Nijmegen/Amsterdam: ITS/SCO-Kohnstamm Instituut.
- Flook, L., Repetti, R., & Ullman, J. (2005). Classroom social experiences as predictors of academic performance. *Developmental Psychology*, 41, 319–327.
- Goldstein, H. (1995). *Multilevel statistical models*. London: Oxford University Press.
- Graham, S., Bellmore, A. D., & Mize, J. (2006). Peer victimization, aggression, and their co-occurrence in middle school: Pathways to adjustment problems. *Journal of Abnormal Child Psychology*, 34, 363–378.
- Graham, S., & Juvonen, J. (1998). Self-blame and peer victimization in middle school: An attributional analysis. *Developmental Psychology*, 34, 587–599.
- Guay, F., Larose, S., & Boivin, M. (2004). Academic self-concept and educational attainment level: A ten-year longitudinal study. *Self and Identity*, 3, 53–68.
- Hanish, L. D., & Guerra, N. G. (2000). The roles of ethnicity and school context in predicting children's victimization by peers. *American Journal of Community Psychology*, 28, 201–223.
- Harter, S. (1988). *Manual for the Self-Perception Profile for Adolescents*. Denver, CO: University of Denver.
- Hawker, D. S. J., & Boulton, M. J. (2000). Twenty years' research on peer victimization and psychosocial maladjustment: A meta-analytic review of cross-sectional studies. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 41, 441–455.
- Juvonen, J., Nishina, A., & Graham, S. (2000). Peer harassment, psychological adjustment, and school functioning in early adolescence. *Journal of Educational Psychology*, 92, 349–359.
- Kapinga, T. L. (2002). *Verantwoording Drempelonderzoek* [Justification threshold research]. Retrieved February 20, 2007, from <http://www.drempelonderzoek.nl/downloads.htm>
- Lopez, C., & DuBois, D. D. (2005). Peer victimization and rejection: Investigation of an integrative model of effects on emotional, behavioral, and academic adjustment in early adolescence. *Journal of Clinical Child and Adolescent Psychology*, 34, 25–36.
- MacKinnon, D. P., Warsi, G., & Dwyer, J. H. (1995). A simulation study of mediated effect measures. *Multivariate Behavioral Research*, 30, 41–62.
- Major, B., Spencer, S., Schmader, T., Wolfe, C., & Crocker, J. (1998). Coping with negative stereotypes about intellectual performance: The role of psychological disengagement. *Personality and Social Psychology Bulletin*, 24, 34–50.
- Marsh, H. W. (1986). The bias of negatively worded items in rating scales for young children: A cognitive-developmental phenomenon. *Developmental Psychology*, 22, 37–49.



- Marsh, H. W. (1987). The big-fish-little-pond-effect on academic self-concept. *Journal of Educational Psychology*, 79, 280-295.
- Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, 76, 397-416.
- McNair, D. M., Lorr, M., & Droppleman, L. F. (1971). *Profile of Mood States manual*. San Diego, CA: Educational and Industrial Testing Service.
- Neblett, E. W., Philip, C. L., Cogburn, C. D., & Sellers, R. M. (2006). African American adolescents' discrimination experiences and academic achievement: Racial socialization as a cultural compensatory and protective factor. *Journal of Black Psychology*, 32, 199-218.
- Rasbash, J., Browne, W., Healy, M., Cameron, B., & Charlton, C. (2004). MLwiN Version 2.0. New York: Multilevel Models Project Institute of Education.
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
- Ryan, R. M., & Deci, E. L. (2000a). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25, 54-67.
- Ryan, R. M., & Deci, E. L. (2000b). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55, 68-78.
- Schwartz, D., Farver, J. M., Chang, L., & Lee-Shin, Y. (2002). Victimization in South Korean children's peer groups. *Journal of Abnormal Child Psychology*, 30, 113-125.
- Schwartz, D., Gorman, A. H., Nakamoto, J., & Toblin, R. L. (2005). Victimization in the peer group and children's academic functioning. *Journal of Educational Psychology*, 97, 425-435.
- Skinner, E. A., Wellborn, J. G., & Connell, J. P. (1990). What it takes to do well in school and whether I've got it: The role of perceived control in children's engagement and school achievement. *Journal of Educational Psychology*, 82, 22-32.
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling*. London: Sage.
- Soenens, B., & Vansteenkiste, M. (2005). Antecedents and outcomes of self-determination in three life domains: The role of parents' and teachers' autonomy support. *Journal of Youth and Adolescence*, 34, 589-604.
- Steele, C. M. (1997). A threat in the air: How stereotypes shape the intellectual identity and performance. *American Psychologist*, 52, 613-629.
- Storch, E. A., Zelman, E., Sweeney, M., Danner, G., & Dove, S. (2002). Overt and relational victimization and psychosocial adjustment in minority preadolescents. *Child Study Journal*, 32, 73-79.
- Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York: Harper Collins.
- Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2006). Self-esteem, academic self-concept, and achievement: How the learning environment moderates the dynamics of self-concept. *Journal of Personality and Social Psychology*, 90, 334-349.
- Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relation between self-beliefs and academic achievement: A meta-analytic review. *Educational Psychologist*, 39, 111-133.
- Verkuyten, M. (1995). Self-esteem, self-concept stability and aspects of ethnic identity among minority and majority youth in the Netherlands. *Journal of Youth and Adolescence*, 24, 155-175.
- Verkuyten, M. (2003). Positive and negative self-esteem among ethnic minority early adolescents: Social and cultural sources and threats. *Journal of Youth and Adolescence*, 32, 267-277.
- Verkuyten, M., Kinket, B., & van der Wielen, C. (1997). The understanding of ethnic discrimination among preadolescents. *Journal of Genetic Psychology*, 138, 97-112.
- Verkuyten, M., & Thijs, J. (2000). *Leren (en) waarden: Discriminatie, zelfbeeld, relaties en leerprestaties in 'witte' en 'zwarte' basisscholen* [Learning and evaluating: Discrimination, self-concept, relations, and achievements in "White" and "Black" primary schools]. Amsterdam: Thela Thesis.
- Verkuyten, M., & Thijs, J. (2001). Peer victimization and self-esteem of ethnic minority group children. *Journal of Community and Applied Social Psychology*, 11, 227-234.
- Verkuyten, M., & Thijs, J. (2002). School satisfaction of elementary school children: The role of performance, peer relations, ethnicity, and gender. *Social Indicators Research*, 59, 203-228.
- Verkuyten, M., & Thijs, J. (2004). Global and ethnic self-esteem in school context: Minority and majority groups in the Netherlands. *Social Indicators Research*, 67, 253-281.
- Verkuyten, M., & Thijs, J. (2006). Ethnic discrimination and global self-worth in early adolescence. *International Journal of Behavioral Development*, 30, 107-116.
- Verkuyten, M., Thijs, J., & Canatan, K. (2001). Achievement motivation and academic performance among Turkish early and young adolescents in the Netherlands. *Genetic, Social, and General Psychology Monographs*, 127, 378-408.
- Wichstrøm, L. (1995). Harter's Self-Perception Profile for Adolescents: Reliability, validity, and evaluation of the question format. *Journal of Personality Assessment*, 65, 100-116.
- Wong, C. A., Eccles, J. S., & Sameroff, A. (2003). The influence of ethnic discrimination and ethnic identification on African American adolescents' school and socioemotional adjustment. *Journal of Personality*, 71, 1197-1232.

Received March 13, 2007

Revision received February 12, 2008

Accepted February 12, 2008 ■

# Engagement and Disaffection in the Classroom: Part of a Larger Motivational Dynamic?

Ellen Skinner  
Portland State University

Carrie Furrer  
NPC Research

Gwen Marchand  
University of Nevada, Las Vegas

Thomas Kindermann  
Portland State University

A study of 805 4th through 7th graders used a model of motivational development to guide the investigation of the internal dynamics of 4 indicators of behavioral and emotional engagement and disaffection and the facilitative effects of teacher support and 3 student self-perceptions (competence, autonomy, and relatedness) on changes in these indicators over the school year. In terms of internal dynamics, emotional components of engagement contributed significantly to changes in their behavioral counterparts; feedback from behavior to changes in emotion were not as consistent. Teacher support and students' self-perceptions (especially autonomy) contributed to changes in behavioral components: Each predicted increases in engagement and decreases in disaffection. Tests of process models revealed that the effects of teacher context were mediated by children's self-perceptions. Taken together, these findings suggest a clear distinction between indicators and facilitators of engagement and begin to articulate the dynamics between emotion and behavior that take place *inside* engagement and the motivational dynamics that take place *outside* of engagement, involving the social context, self-systems, and engagement itself.

**Keywords:** academic engagement, disaffection, achievement motivation, classroom participation, emotional engagement

Over the past 10 years, research has converged on the construct of academic engagement as a key contributor to children's school success (Fredricks, Blumenfeld, & Parks, 2004). In the short term, engagement predicts students' learning, grades, and achievement test scores; over the long term, it predicts patterns of attendance, retention, graduation, and academic resilience (Connell, Spencer, & Aber, 1994; Finn & Rock, 1997; Jimerson, Campos, & Greif, 2003; Sinclair, Christenson, Lehr, & Anderson, 2003; Skinner, Zimmer-Gembeck, & Connell, 1998). Studies have also suggested that academic engage-

ment serves as a protective factor against risky activities (O'Farrell & Morrison, 2003) such as substance abuse, risky sexual behavior, and delinquency.

Thus, students who are engaged in school are both more successful academically and more likely to avoid the pitfalls of adolescence. Unfortunately, however, research has also documented a steady decline in students' engagement with schooling, including their interest, enthusiasm, and intrinsic motivation for learning in school, beginning in kindergarten and continuing until they complete high school (or drop out), with notable losses during the transitions to middle school and high school; the erosion of engagement is especially severe for boys and for children from ethnic and racial minority and low socioeconomic status groups (for a review, see Wigfield, Eccles, Schiefele, Roeser, & Davis-Kean, 2006).

---

Ellen Skinner and Thomas Kindermann, Department of Psychology, Portland State University; Carrie Furrer, NPC Research, Portland, OR; Gwen Marchand, Department of Educational Psychology, University of Nevada, Las Vegas.

We express our gratitude to the Motivation Research Group, especially James Connell, Edward Deci, and Richard Ryan. We would especially like to express our appreciation to James Wellborn for his earlier work on conceptualization and measures of engagement. In terms of the research project, we thank the Brockport School District and its superintendent, principals, teachers, students, and parents for their generous participation. The hard work and good spirits of the research team members are gratefully acknowledged, including Jeff Altman, Michael Belmont, Helen Dorsett, Jennifer Herman, Marianne Miserandino, Brian Patrick, Cara Regan, Hayley Sherwood, and Peter Usinger. We acknowledge support from the W. T. Grant Foundation, from National Institute of Child Health and Human Development Research Grant No. HD19914, and from National Institutes of Mental Health Training Grant No. 527594.

Correspondence concerning this article should be addressed to Ellen A. Skinner, Psychology Department, P.O. Box 751, Portland State University, Portland, OR 97221. E-mail: skinnere@pdx.edu

## Motivational Dynamics of Classroom Engagement and Disaffection

Of great interest to researchers and practitioners are the dynamics underlying these declines. *Dynamics* refers to the internal and external causal feedback loops that serve to promote or undermine the quality of children's engagement in school over time. In general, these dynamics seem to be amplifying, in that children who start out motivationally rich maintain their engagement as the year(s) progress, whereas children who start out motivationally poor tend to become even more disengaged over time (Skinner, Kindermann, Connell, & Wellborn, in press).

Some of these dynamics involve personal motivational resources, such as perceived control: For example, children who start



off confident in their capacities engage with learning tasks in ways that lead to more success, thus reinforcing their initial optimism, whereas children low in efficacy tend to avoid challenges or engage in tasks so half-heartedly that they do not succeed, thereby cementing their initial self-doubts (e.g., Schmitz & Skinner, 1993). Other dynamics involve the teacher: Children who are more engaged receive subsequently more teacher involvement, whereas disaffected children are more likely to find that teachers increasingly withdraw their support or become more controlling over time (e.g., Skinner & Belmont, 1993). It is also possible that some of these dynamics are internal to engagement itself. If multiple components of engagement can be distinguished, they may form their own feedback loops. For example, children who are bored may exert less effort and stop paying attention to the teacher, thus becoming even more bored over time.

Purposes of the Current Study

The current study makes two key contributions to emerging work on motivational dynamics. First, we explore the internal dynamics of engagement by examining how different components of engagement shape each other over time. Second, we explore the larger motivational dynamics of which engagement is a part by examining how contextual and personal factors contribute to changes in engagement itself. Both the components of classroom engagement and the set of facilitators hypothesized to promote them are derived from a larger motivational model, the self-system model of motivational development (SSMMD), which can be used to explain the interpersonal and psychological processes by which engagement is promoted or undermined in the classroom (Connell & Wellborn, 1991; Deci & Ryan, 1985; Skinner & Wellborn, 1997).

Conceptualization of Engagement

Operationalizations of engagement have been offered from a variety of theoretical and practice approaches, leading reviewers to conclude that it is a metaconstruct encompassing multiple dimensions of attraction to or involvement in school (Fredricks et al., 2004). However, two important areas of confusion remain. The first focuses on the distinction between indicators versus facilitators of engagement (Sinclair et al., 2003). *Indicators* refer to the features that belong inside the construct of engagement proper, whereas *facilitators* are the causal factors (outside of the construct) that are hypothesized to influence engagement. Explanatory research and intervention efforts require a clear demarcation between these two. If, for example, conceptualizations posit that support provided by teachers is part of engagement itself (i.e., an indicator) as opposed to a contextual factor that contributes to engagement (i.e., a facilitator), studies that aggregate these features into a metaconstruct can never explore how teacher context shapes children’s engagement. To empirically examine how potential antecedents influence engagement, it is necessary to conceptually unpack indicators from facilitators.

The second issue requiring clarification centers on the number and nature of dimensions within engagement itself: how many should be distinguished and whether they have their own internal dynamics. Some reviewers have suggested that it is useful to distinguish affective, behavioral, and cognitive forms (e.g.,

Fredricks et al., 2004), but little agreement exists as to what these components entail. Others suggest that “good news” should be differentiated from “bad news” features, based on the argument that alienation and disaffection likely reflect more than a lack of engagement (Jimerson et al., 2003). Clear definitions, sound assessments, and evidence of multiple dimensions are required to answer questions about the components of engagement.

Indicators of Classroom Engagement

In this study, we used a motivational conceptualization of engagement versus disaffection, which focuses on students’ active participation in academic activities in the classroom (Pierson & Connell, 1992; Ryan, 2000; Skinner et al., 1998; Wentzel, 1993). The underlying assumption is that high-quality learning is the result of behaviors and emotions, such as exertion, persistence, interest, and enjoyment, that reflect a motivation to master the academic material. As depicted in Figure 1, this conceptualization incorporates behavioral and emotional dimensions, as well as a specific treatment of negative engagement referred to as *disaffection* (Connell & Wellborn, 1991; Skinner, Kindermann, & Furrer, in press; Wellborn, 1991).

The behavioral dimension of engagement includes students’ effort, attention, and persistence during the initiation and execution of learning activities. The emotional dimension of engagement focuses on states that are germane to students’ emotional involvement during learning activities such as enthusiasm, interest, and enjoyment (Meyer & Turner, 2002). Engagement itself combines behavioral and emotional dimensions and refers to active, goal-directed, flexible, constructive, persistent, focused, emotionally positive interactions with the social and physical environments (in this case, academic activities). Consistent with the SSMMD, this kind of engagement has been found to be a strong predictor of

	ENGAGEMENT	DISAFFECTION
B E H A V I O R	<b>Behavioral Engagement</b>	<b>Behavioral Disaffection</b>
	Action initiation	Passivity
	Effort, Exertion	Giving up
	Attempts, Persistence	Withdrawal
	Intensity	Inattentive
	Attention, Concentration	Distracted
	Absorption	Mentally disengaged
E M O T I O N	<b>Emotional Engagement</b>	<b>Emotional Disaffection</b>
	Enthusiasm	Boredom
	Interest	Disinterest
	Enjoyment	Frustration/anger
	Satisfaction	Sadness
	Pride	Worry/anxiety
	Vitality	Shame
	Zest	Self-blame

Figure 1. A motivational conceptualization of engagement and disaffection in the classroom.

student learning, grades, achievement, and school retention (e.g., Connell, Halpern-Fisher, Clifford, Crichlow, & Usinger, 1995; Connell et al., 1994; Skinner, Wellborn, & Connell, 1990).

Disaffection, which signifies more than the absence of engagement, refers to the occurrence of behaviors and emotions that reflect maladaptive motivational states. Disaffection has both a behavioral component, including passivity and withdrawal from participation in learning activities, and an emotional component, including boredom, anxiety, and frustration in the classroom. Disaffection has been found to be a strong predictor of poor grades, low achievement test scores, and eventual drop out (e.g., Connell et al., 1994, 1995; Skinner et al., 1990).

Recent psychometric work has suggested that these four markers of classroom engagement, namely, behavioral and emotional engagement and behavioral and emotional disaffection, are structurally distinguishable (Furrer, Skinner, Marchand, & Kindermann, 2006; Skinner, Kindermann, & Furrer, in press). More specifically, item sets tapping each of the four dimensions were used to directly compare structural models with different numbers of dimensions. Although the dimensions were closely related, a four-factor model showed a significantly better fit to both student- and teacher-report data than either one- or two-factor models. This pattern was found for elementary, middle, and high school students. Evidence of additional (hierarchical) multidimensionality was found for emotional disaffection: Item sets tapping boredom, frustration, and anxiety were better represented by three dimensions than by a single dimension.

### *Internal Dynamics of Engagement and Disaffection*

Although researchers have suggested the importance of deconstructing engagement and examining how the parts work together before combining them into an aggregate or metaconstruct (e.g., Fredricks et al., 2004), theories depicting the internal dynamics of engagement, that is, how the components of engagement mutually influence each other over time, have not been fully articulated. The baseline proposition is that there are none: Behavioral and emotional engagement in the classroom are tightly coupled, largely interindividually stable, and shaped in the same ways by outside factors, without influencing each other. However, when internal dynamics are mentioned, it is usually with the idea that emotions fuel behaviors in the classroom. For example, self-determination theory (Deci & Ryan, 1985) and theories of effectance motivation (Harter, 1978) suggest that it is engaged emotions, such as interest and enthusiasm, that fuel engaged behaviors, such as effort and persistence.

Emotions may also play a leading role in the dynamics of how students lose engagement and become disaffected as the school years progress (e.g., Finn, Pannozzo, & Voelkl, 1995; Roeser, Strobel, & Quihuis, 2002). That is, if students become bored, frustrated, or anxious about schoolwork, this likely undermines their behavioral participation in academic activities. Because the disaffected emotions can be differentiated, it is possible that they may have different effects on behavior. For example, boredom, a relatively passive emotion, might result in losses in behavioral engagement, but perhaps not lead to more overt behavioral disaffection. However, an emotion like frustration might be more strongly linked to active behavioral disaffection in the classroom. Hence, on the basis of recent research that distinguishes four

indicators, the first contribution of the present study was to examine how these components work together over time.

### *Facilitators of Classroom Engagement*

The second goal of the current study was to examine the processes through which an engaged dynamic is created and maintained in the classroom. Hypotheses were drawn from the SSMMD, which focuses on engagement but depicts a larger motivational dynamic (Connell & Wellborn, 1991; Deci & Ryan, 1985). The SSMMD includes four basic higher order constructs: context, self, action, and outcomes (see Figure 2). The general hypothesis, supported by accumulating empirical evidence, is that a more supportive classroom context promotes positive self-perceptions, which in turn fuel engagement in the classroom; conversely, a less supportive classroom context undermines self-perceptions, which then feed disaffection with learning. Empirical support for each of the links posited by the model is described briefly below, with a special focus on differential predictions for behavioral versus emotional engagement.

### *Self-System Processes and Classroom Engagement Versus Disaffection*

Within the SSMMD, self-system processes (SSPs) are defined as relatively durable personal resources (or liabilities) that individuals construct over time in response to interactions with the social context; they are organized around people's basic needs for competence, autonomy, and relatedness. SSPs, within this framework, are proximal predictors of engagement and disaffection. Hence, the SSMMD holds that beliefs about the self can be distinguished from engagement. This distinction is important because in the larger literature on academic engagement, a group of constructs with the common theme of interpersonal relationships (e.g., school attachment, school bonding, and school belonging) has been classified as a dimension of engagement itself (Jimerson et al., 2003). The SSMMD pulls the interpersonal relationship piece out of the definition of engagement and establishes SSPs as facilitators rather than indicators of engagement. Each of the three SSPs has a long history of study under a variety of labels.

*Competence.* Competence is perhaps the most frequently studied self-perception in the academic domain (Wigfield et al., 2006). According to the SSMMD, individuals are born with the need to experience themselves as effective in their interactions with the environment (Elliot & Dweck, 2005; White, 1959), and the extent to which they feel this sense of mastery is related to the quality of their engagement in that domain. Perceptions of self-efficacy, ability, academic competence, and control are robust predictors of children's effort and persistence in school and of their emotional reactions to success and failure (see Bandura, 1997; Dweck, 1999; Elliot & Dweck, 2005; Harter, 1982; Skinner, 1995; Stipek, 2002; Weiner, 2005; Wigfield et al., 2006).

*Autonomy.* Following self-determination theory (Deci & Ryan, 1985), the model holds that individuals are born with the need to express their genuine preferences and act in congruence with their true selves; the extent to which individuals experience autonomy in a particular domain is related to the quality of their engagement in that domain. Studies have generally shown that students with a greater sense of autonomy in school settings have



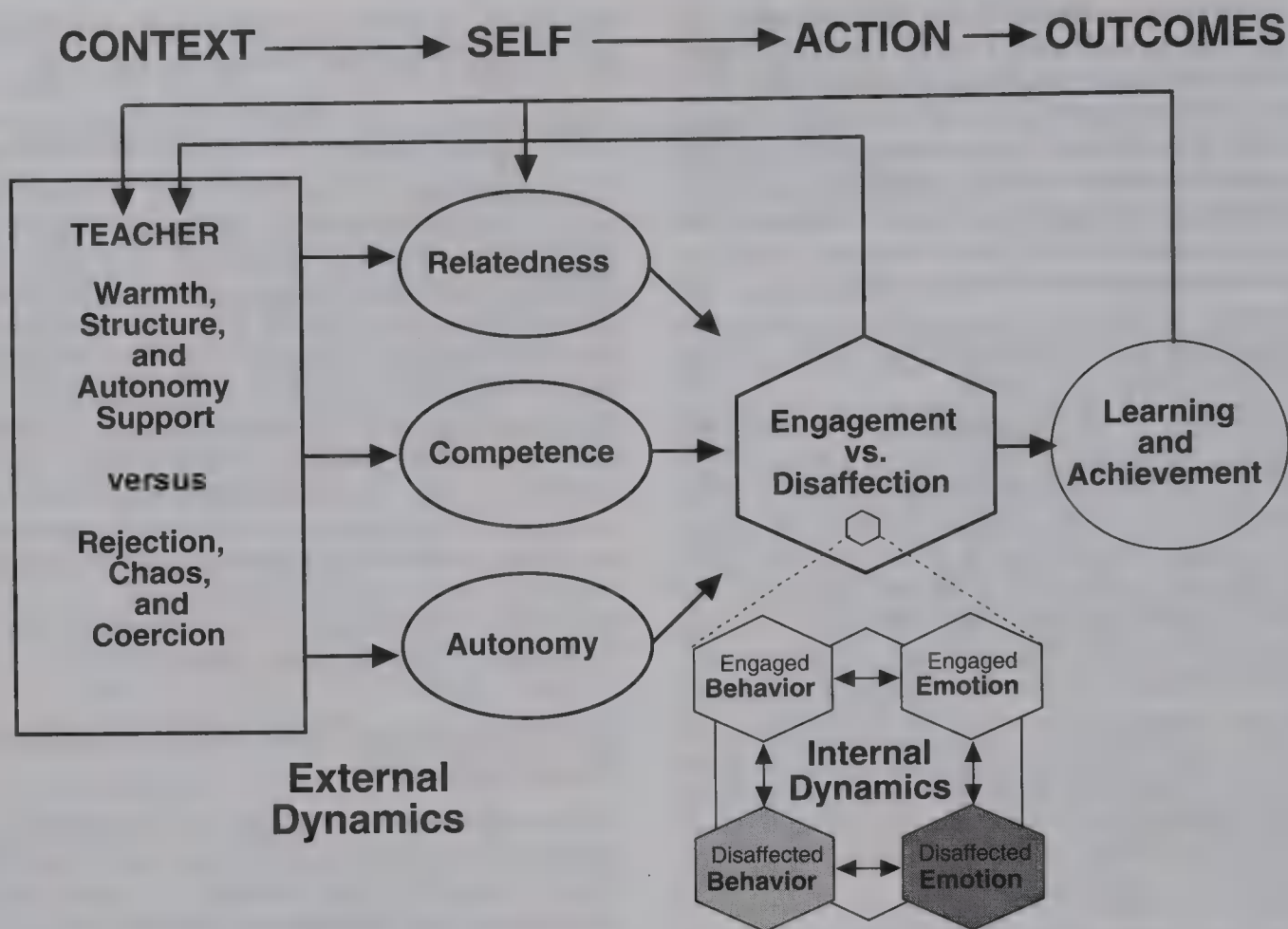


Figure 2. The self-system model of motivational development, including dynamics internal to engagement and external dynamics that incorporate engagement.

better academic outcomes such as classroom engagement, persistence, achievement, and learning (e.g., Grolnick & Ryan, 1987; Hardre & Reeve, 2003; Miserandino, 1996; Patrick, Skinner, & Connell, 1993; Vallerand, Fortier, & Guay, 1997).

**Relatedness.** Relatedness tends to be overlooked as a self-perception in the academic domain. From a motivational perspective, the basic concept is that individuals are born with an innate desire to connect to others (Ainsworth, Blehar, Waters, & Wall, 1978; Bowlby, 1969/1973; Baumeister & Leary, 1995) and that the extent to which they feel that they belong in a particular enterprise is associated with the quality of their engagement in the activities of that enterprise. Research has documented a link between a sense of belonging in school and multiple indicators of academic motivation and adjustment, especially emotional engagement (Anderman, 1999; Anderman & Anderman, 1999; Battistich, Solomon, Watson, & Schaps, 1995; Eccles & Midgley, 1989; Furrer & Skinner, 2003; Goodenow, 1993; Kuperminc, Blatt, Shahar, Henrich, & Leadbetter, 2004; Lynch & Cicchetti, 1997; Roeser, Midgley, & Urdan, 1996; Ryan, Stiller, & Lynch, 1994; Wentzel, 1997, 1998, 1999).

#### *Contextual Supports Shape Classroom Engagement Versus Disaffection*

The SSMMD holds that contextual features are critical in promoting motivation in the classroom (Connell & Wellborn, 1991; Deci & Ryan, 1985; Stipek, 2002; Weiner, 1990). Although students' moti-

vation is shaped by multiple social partners (Wentzel, 1998), the current study targets supportive interactions with teachers. Support for motivation includes pedagogical caring (Wentzel, 1997) as well as autonomy-supportive instruction (e.g., relevance and giving students choices; Guthrie & Davis, 2003; Reeve, Bolt, & Cai, 1999; Reeve, Jang, Carrell, Jeon, & Barch, 2004) and optimal structure (e.g., predictability and responsiveness; Skinner et al., 1998). The quality of student-teacher relationships, in the form of caring, supportive alliances, is a key predictor of academic engagement, effort in the classroom, school liking, and achievement expectancies (Birch & Ladd, 1997, 1998; Goodenow, 1993; Murdock, 1999; Murray & Greenberg, 2000; Ryan & Powelson, 1991).

*SSPs mediate the relationship between context and classroom engagement.* Finally, the SSMMD also posits a specific mediated pathway in which features of the context influence how individuals feel about themselves (i.e., SSPs), which in turn predicts whether they will be engaged or disaffected in that context. Fredricks et al. (2004), in their review of the construct of engagement, noted that "few scholars include measures of context, needs, and engagement in the same study" (p. 80). One study that did test mediated pathways between context and engagement showed that positive student-teacher relationships were connected to a sense of school belonging, which in turn predicted positive affect in school (Roeser et al., 1996). The current study addresses this empirical gap by testing a mediated pathway from teacher support to self-perceptions to engagement versus disaffection.

### Summary of Hypotheses

In sum, on the basis of a clear distinction between indicators and facilitators of engagement, we aimed to examine some of the dynamics underlying the general decline in motivation during late elementary and early middle school. Using information collected from fourth through seventh graders at the beginning and end of the same school year, we first attempted to replicate the pattern of between-year differences and within-year decrements in engagement that suggest general motivational losses over the transition to middle school. Second, we examined the dynamics between emotion and behavior that take place inside engagement, testing the hypothesis that not only would behavior predict changes in emotion, but that emotion would be an even stronger predictor of changes in behavior. We further decomposed emotional disaffection into a set of multiple indicators (frustration, boredom, and anxiety) to examine whether they played a differential role in predicting changes in the other facets of engagement.

Third, we examined the motivational dynamics that take place outside of engagement, involving the social context, self-systems, and engagement itself. We predicted that children's SSPs would contribute to changes in their engagement over the school year, with perceived competence being perhaps the biggest predictor of changes in behavior and relatedness and autonomy being stronger predictors of changes in emotion. Moreover, we expected teacher support to predict changes in student engagement and disaffection, with SSPs representing an important pathway through which teachers' support would be connected to engagement. Although we expected mean-level differences in engagement according to gender and grade (favoring girls and younger children), we nevertheless hypothesized that the dynamics of engagement would not differ among these groups. In sum, the present study has the potential to help organize conceptualizations of the complex multidimensional construct of engagement and to contribute to ongoing research that examines the social and personal factors that shape its development over time.

### Method

#### Participants

Data from 805 children (195 fourth graders, 131 fifth graders, 290 sixth graders, and 189 seventh graders approximately equally divided by gender) who had participated in a 4-year longitudinal study on children's motivation in school were used from two measurement points (fall and spring of Year 4). Students and their 53 teachers, drawn from the only public elementary and middle schools in a rural-suburban school district in upstate New York, were predominantly Caucasian, with only about 5% of the students identifying themselves as non-White. Student socioeconomic status, as determined by parents' level of education and occupation, ranged between working to middle class. (See Skinner et al., 1998, for details.) From the 1,242 children who provided any data during Year 4, we selected a subset of 805 children who were missing less than 5% of their data. For this subsample, missing data were imputed with SPSS 11.5 using maximum likelihood estimation with an estimation maximization algorithm.

#### Procedures

Students completed self-report questionnaires administered by trained interviewers in three 45-min sessions. In their normal class-

rooms, students marked questionnaire items as they were read aloud by one interviewer; a second interviewer monitored understanding and answered questions. Teachers were not present; for the most part, they filled out their questionnaires while students were being tested. Questionnaires were administered in October and again in May.

### Measures

Students reported on their engagement versus disaffection in the classroom, their sense of perceived competence and control in the academic domain, autonomy in the classroom, and relatedness to their teacher and their impressions of the support they received from teachers. Teachers reported on the support they provided to each student. If a student had multiple teachers, information was provided from the teacher who indicated that he or she knew the student the best. Each scale was made up of positively and negatively worded items. Composite scores were determined by calculating the average of the positive and negative items, then reverse coding the negative subscale and averaging it with the positive subscale. Resulting scores ranged from 1 to 4, with higher scores indicating more of the respective construct.

*Behavioral and emotional engagement and disaffection.* Students reported on their own engagement versus disaffection in the classroom using a measure developed to tap their behavioral and emotional participation in (or withdrawal from) learning activities in the classroom (Skinner et al., 1990, 1998; Wellborn, 1991). Behavioral engagement was assessed using 5 items that tapped students' effort, attention, and persistence while initiating and participating in learning activities ( $\alpha = .71$  in fall;  $\alpha = .72$  in spring). Behavioral disaffection was assessed using 5 items that tapped students' lack of effort and withdrawal from learning activities ( $\alpha = .65$  in fall;  $\alpha = .70$  in spring). Emotional engagement was measured using 6 items that tapped emotions indicating students' motivated participation during learning activities ( $\alpha = .83$  in fall;  $\alpha = .84$  in spring). Emotional disaffection was measured using 10 items that tapped emotions indicating students' motivated withdrawal or alienation during learning activities. Items were averaged according to the specific emotions (boredom, anxiety, and frustration) and then combined for a summary score ( $\alpha = .84$  in fall;  $\alpha = .84$  in spring). Items from the current version of the student-report measure are presented in the Appendix. (See also Skinner, Kindermann, & Furrer, in press.)

*Perceived competence and control.* The Control Beliefs subscale of the Student Perceptions of Control Questionnaire (Skinner, Chapman, & Baltes, 1983, 1988; Skinner et al., 1990) was used to tap children's perceived competence. The Control Beliefs subscale consists of six items tapping students' generalized expectancies about the extent to which they can achieve success and avoid failure in school. Examples of items are "I can do well in school if I want to" and "I can't get good grades, no matter what I do" (reverse coded; Skinner et al., 1990, 1998). Items were averaged to form a summary score ( $\alpha = .74$  in fall;  $\alpha = .73$  in spring).

*Autonomy orientation.* The Autonomy Scale, used to assess academic autonomy (Ryan & Connell, 1989), is composed of 17 items that tap whether children engage in activities because they feel pressured or because they desire understanding and enjoy the task, divided into four subscales: (a) External self-regulation refers to doing work because of rules or fear of punishment ("Why do I do my homework? Because I'll get in trouble if I don't"); (b)



introjected self-regulation refers to doing work because one "should" and to avoid negative emotions ("Because I'll feel really bad about myself if I don't do well"); (c) identified self-regulation refers to doing work to understand more ("Because I think class-work is important for my learning"); and (d) intrinsic self-regulation refers to doing work because it is enjoyable ("Because it's fun"). Subscales for external and introjected self-regulation were reverse coded and then averaged with subscales for identified and intrinsic self-regulation to form a summary score ( $\alpha = .81$  in fall;  $\alpha = .81$  in spring).

**Sense of relatedness.** Students completed a four-item self-report scale tapping a sense of belonging or relatedness to their teachers (Furrer & Skinner, 2003). For each item, the stem was "When I'm with my teacher" and the items were "I feel accepted," "I feel like someone special," "I feel ignored" (reverse coded), and "I feel unimportant" (reverse coded). The items were averaged to form a summary score ( $\alpha = .82$  in fall;  $\alpha = .84$  in spring).

**Teacher support: Student report.** Students reported on the involvement, structure, and autonomy support they experienced from their teachers (Skinner & Belmont, 1993). Nine items tapped involvement versus hostility, including warmth-affect, dedication of resources, knowledge about the student, and dependability versus hostility and neglect (reverse coded). Example items include "My teacher likes me" and "My teacher doesn't seem to enjoy having me in her class" (reverse coded). Twenty-one items measured provision of structure, including clarity of expectations and contingency, versus chaos (reverse coded). Example items are "My teacher shows me how to solve problems for myself" and "My teacher doesn't make clear what she expects of me in class" (reverse coded). Eighteen items tapped autonomy support versus coercion, including teacher provision of choice, relevance, and respect versus controlling behavior (reverse coded). Example items are "My teacher gives me a lot of choices about how I do my schoolwork" and "It seems like my teacher is always telling me what to do" (reverse coded). Scales were averaged to form a Teacher Support scale ( $\alpha = .96$  in fall;  $\alpha = .96$  in spring).

**Teacher support: Teacher report.** Teachers reported on the level of involvement, structure, and autonomy support they pro-

vided to each child (Skinner & Belmont, 1993). Fourteen items tapped involvement versus hostility, including warmth-affect, dedication of resources, knowledge about the student, and dependability versus hostility and neglect (reverse coded). Example items include "This student is easy to like" and "Teaching this student is not very enjoyable" (reverse coded). Five items measured provision of structure, including clarity of expectations and contingency, versus chaos (reverse coded). Example items are "I try to be clear with this student about what I expect of him/her in class" and "I find it hard to be consistent with this student" (reverse coded). Twelve items tapped autonomy support versus coercion, including teacher provision of choice, relevance, and respect versus controlling behavior (reverse coded). Example items are "I let this student make a lot of his/her own decisions regarding schoolwork" and "When it comes to assignments, I'm always having to tell this student what to do" (reverse coded). Scales were averaged to form a Teacher Support scale ( $\alpha = .95$  in fall;  $\alpha = .95$  in spring).

## Results

Initial analyses examined descriptive information. As can be seen in Table 1, indicators of engagement versus disaffection suggested that averaged over fall and spring, students were moderately engaged ( $M = 3.21$ ,  $SD = 0.50$ , for behavioral and emotional engagement combined) and not particularly disaffected ( $M = 2.07$ ,  $SD = 0.57$ , for behavioral and emotional disaffection combined). In terms of facilitators, children reported relatively high levels of all three self-systems averaged over fall and spring, although they reported higher competence ( $M = 3.44$ ,  $SD = 0.49$ ) than relatedness ( $M = 3.03$ ,  $SD = 0.69$ ) and higher relatedness than autonomy ( $M = 2.58$ ,  $SD = 0.45$ ). In addition, teachers were perceived as supportive ( $M = 2.95$ ,  $SD = 0.52$ ) and themselves reported providing relatively high support ( $M = 3.15$ ,  $SD = 0.38$ ).

### Grade Differences and Changes in Engagement and Disaffection

To determine whether we could replicate the pattern of between-grade differences and within-grade declines indicating losses in

Table 1  
Descriptive Statistics for Indicators and Facilitators of Engagement

Construct	Fall	<i>M</i> ( <i>SD</i> )	Spring	<i>M</i> ( <i>SD</i> )	<i>t</i>	Fall to spring <i>r</i>
Indicators of engagement						
Behavioral engagement	3.45	(0.47)	3.33	(0.50)	-7.29***	.57***
Behavioral disaffection	1.94	(0.61)	1.99	(0.61)	-2.44*	.67***
Emotional engagement	3.07	(0.62)	2.99	(0.65)	4.08***	.63***
Emotional disaffection	2.14	(0.62)	2.22	(0.63)	-4.47***	.65***
Bored	2.28	(0.91)	2.42	(0.91)	-4.92***	.59***
Anxious	2.00	(0.70)	2.06	(0.72)	-2.82**	.62***
Frustrated	2.26	(0.68)	2.32	(0.69)	-2.47*	.53***
Facilitators of engagement						
Perceived control	3.48	(0.53)	3.40	(0.55)	4.58***	.61***
Autonomy orientation	2.61	(0.49)	2.54	(0.49)	5.11***	.69***
Sense of relatedness	3.08	(0.77)	2.97	(0.80)	4.00***	.55***
Teacher support						
Student report	2.99	(0.56)	2.91	(0.55)	6.67***	.76***
Teacher report	3.19	(0.42)	3.12	(0.37)	8.77***	.82***

Note.  $N = 805$ . All scales ranged from 1 (not at all true for me/this student) to 4 (very true for me/this student).

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

motivation over the transition to middle school, we first examined each indicator and facilitator as a function of grade; see Table 2 for means and standard deviations (averaged across the two time points). As is typical (Fredricks et al., 2004; Wigfield et al., 2006), starting in sixth grade (during the transition to middle school), older children showed lower levels of engagement and higher levels of disaffection. Profile analyses, in which the four indicators of engagement were used as within-subject dependent variables, indicated that the profiles of engagement differed by grade,  $F(12, 2111) = 8.64, p < .001$ . As depicted in Figure 3, children in fourth and fifth grades showed a profile in which engagement was high and disaffection was low. However, after the transition to middle school, students showed lower levels of engagement, especially emotional engagement, and higher levels of disaffection, especially emotional disaffection, and this trend continued to worsen in seventh grade. In addition, older children showed lower levels of self-systems and (both teacher- and student-reported) teacher support, with the most noticeable differences starting in middle school between fifth and sixth and between sixth and seventh grades.

Consistent with these differences between grades, there was also a slight worsening in both indicators and facilitators of engagement within grades, in that engagement decreased and disaffection increased from fall to spring, and self-systems and teacher support declined (see Table 1). At the same time, interindividual stability was high: Cross-year correlations for student-report variables averaged .68 (all  $ps < .001$ ). This pattern of findings, namely high interindividual stability combined with motivational declines that are experienced differentially depending on initial levels of engagement, are consistent with the typical amplifying pattern of loss whose underlying dynamics this study was designed to illuminate.

**Gender differences.** Table 2 also presents the means and standard deviations for each indicator and facilitator (averaged across the two time points) broken down by gender. As is typical (Fredricks et al., 2004; Wigfield et al., 2006), mean-level differ-

ences favored girls, although there were no differences in emotional disaffection or any of the differentiated disaffected emotions. In each of the following analyses, we also examined interactions with grade and gender. In no case were they significant, suggesting that despite mean-level differences favoring younger children and girls, the dynamics of engagement played out in a similar manner across fourth through seventh grades and for the boys and girls in this sample.

### *Internal Dynamics Among Indicators of Engagement Versus Disaffection*

The second set of analyses examined the internal dynamics of engagement and disaffection. As can be seen in Table 3, the four indicators showed the expected concurrent interrelations. (See Skinner, Kindermann, & Furrer, in press, for supporting evidence from confirmatory structural analyses.) To test hypotheses about predictors of change over the school year, we calculated multiple regressions in which we controlled for the dependent variable in the fall before examining whether the independent variable in fall was a significant predictor of the dependent variable in spring.

**Emotions as predictors of changes in behavioral engagement.** Figure 4 depicts the regressions that examined whether emotional engagement predicted changes in behavioral engagement and behavioral disaffection. As expected, despite the high stabilities of the dependent variables, emotional engagement in the fall significantly predicted improvements in behavioral engagement and declines in behavioral disaffection from fall to spring.

The top panel of Figure 4 also depicts the regressions testing whether emotional disaffection predicted changes in behavioral engagement and disaffection. As can be seen, emotional disaffection in the fall contributed significantly to increases in behavioral disaffection and decreases in behavioral engagement from fall to spring. Regressions examining the differentiated disaffected emo-

Table 2  
*Descriptive Statistics by Gender and Grade Averaged Across the School Year*

Construct	<i>M (SD)</i>		Gender <i>t</i>	<i>M (SD)</i>					Grade <i>F</i>
	Boys ( <i>n</i> = 393)	Girls ( <i>n</i> = 412)		Grade 4 ( <i>n</i> = 195)	Grade 5 ( <i>n</i> = 131)	Grade 6 ( <i>n</i> = 290)	Grade 7 ( <i>n</i> = 189)		
Indicators of engagement									
Behavioral engagement	3.31 (0.52)	3.47 (0.44)	-5.16***	3.57 (0.40) <sub>a</sub>	3.54 (0.43) <sub>a</sub>	3.36 (0.49)	3.15 (0.48)	41.19***	
Behavioral disaffection	2.06 (0.61)	1.87 (0.06)	4.88***	1.92 (0.56) <sub>b, c</sub>	1.79 (0.64) <sub>b</sub>	1.97 (0.61) <sub>c</sub>	2.12 (0.60)	9.63***	
Emotional engagement	2.96 (0.66)	3.09 (0.60)	-3.14**	3.22 (0.57) <sub>d</sub>	3.17 (0.64) <sub>d</sub>	2.98 (0.63)	2.81 (0.63)	20.91***	
Emotional disaffection	2.19 (0.63)	2.17 (0.62)	0.45	2.00 (0.61) <sub>e</sub>	2.04 (0.62) <sub>e</sub>	2.28 (0.62) <sub>f</sub>	2.32 (0.59) <sub>f</sub>	16.92***	
Bored	2.38 (0.90)	2.31 (0.92)	1.27	2.03 (0.88) <sub>g</sub>	2.10 (0.88) <sub>g</sub>	2.47 (0.89)	2.67 (0.84)	28.36***	
Anxious	2.01 (0.70)	2.05 (0.72)	-0.86	1.98 (0.71) <sub>h, i, j</sub>	1.91 (0.70) <sub>h, k</sub>	2.10 (0.70) <sub>i, l</sub>	2.07 (0.71) <sub>j, k, l</sub>	3.70*	
Frustrated	2.32 (0.70)	2.27 (0.67)	1.16	2.09 (0.67) <sub>m</sub>	2.21 (0.65) <sub>m</sub>	2.38 (0.70) <sub>n</sub>	2.42 (0.66) <sub>n</sub>	13.45***	
Facilitators of engagement									
Perceived control	3.37 (0.54)	3.50 (0.53)	-3.64***	3.48 (0.52) <sub>o, p</sub>	3.55 (0.48) <sub>o, q</sub>	3.44 (0.55) <sub>p, q, r</sub>	3.32 (0.57) <sub>r</sub>	6.71***	
Autonomy orientation	2.52 (0.46)	2.62 (0.51)	-3.15**	2.77 (0.49) <sub>s</sub>	2.70 (0.49) <sub>s</sub>	2.53 (0.47)	2.35 (0.42)	36.29***	
Sense of relatedness	2.91 (0.79)	3.13 (0.76)	-4.50***	3.14 (0.78) <sub>t</sub>	3.24 (0.80) <sub>t</sub>	2.96 (0.76) <sub>u</sub>	2.85 (0.75) <sub>u</sub>	11.54***	
Teacher support									
Student report	2.88 (0.55)	3.02 (0.56)	-3.79***	3.13 (0.53) <sub>v</sub>	3.14 (0.56) <sub>v</sub>	2.90 (0.52)	2.72 (0.54)	29.60***	
Teacher report	3.10 (0.40)	3.20 (0.38)	-3.92***	3.13 (0.34) <sub>w, x</sub>	3.32 (0.42)	3.11 (0.37) <sub>w, y</sub>	3.14 (0.44) <sub>x, y</sub>	10.44***	

*Note.* Grade means with the same subscripts did not differ from each other at least at the  $p < .05$  level as identified with post hoc tests using the Bonferroni correction.

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .



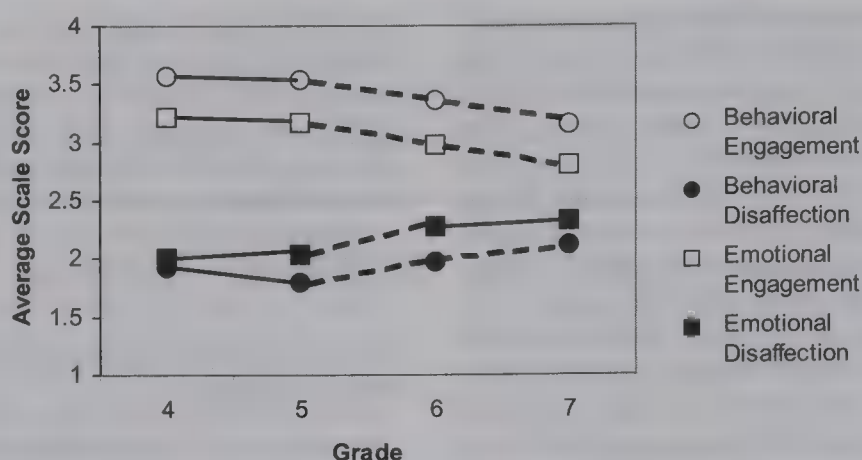


Figure 3. Grade differences in the four components of behavioral and emotional engagement and disaffection.  
 Note. Solid black lines indicate adjacent grade levels that do not differ significantly at least at the  $p < .05$  level.  
 Dotted black lines indicate adjacent grade levels that do differ significantly at least at the  $p < .05$  level.

tions (bored, anxious, or frustrated) revealed they were significant predictors. Each predicted declines in behavioral engagement (bored,  $\beta = -0.23$ ; anxious,  $\beta = -0.10$ ; and frustrated,  $\beta = -0.12$ ,  $ps < .001$ ) and increases in behavioral disaffection (bored,  $\beta = 0.18$ ; anxious,  $\beta = 0.08$ ; and frustrated,  $\beta = 0.09$ ,  $ps < .01$ ).

**Behavior as a predictor of changes in emotional engagement.** The bottom panel of Figure 4 presents the results of regressions examining whether behavioral engagement in the classroom predicts changes in emotional engagement and disaffection over the school year. As can be seen, behavioral engagement in the fall was a significant predictor of increases in emotional engagement. However, it was not a significant predictor of declines in emotional disaffection. Regressions examining behavioral engagement as a predictor of changes in the differentiated negative emotions revealed that it was a significant predictor of declines in boredom ( $\beta = -0.08$ ,  $p < .01$ ) but not in anxiety or frustration.

The bottom panel of Figure 4 also presents the results of the regressions examining whether behavioral disaffection in the classroom contributes to changes in emotion. As expected, despite the high stabilities of the dependent variables, behavioral disaffection in the fall significantly predicted declines in emotional engagement and increases in emotional disaffection from fall to spring. Regressions examining behavioral disaffection as a predictor of changes in the differentiated negative emotions revealed that

it was a significant predictor of increases in each (bored,  $\beta = 0.12$ ; anxious,  $\beta = 0.14$ ; and frustrated,  $\beta = 0.12$ ,  $ps < .001$ ).

**Feedforward and feedback effects.** To determine whether, as predicted, emotion had stronger feedforward effects on behavior compared with the feedback effects of behavior on emotion, analyses directly compared the coefficients for each pair of predictors and outcomes by subtracting the unstandardized regression coefficients and dividing them by the pooled standard error. Only one pair was significantly different: The regression coefficient depicting the feedforward effect of emotional engagement on changes in behavioral engagement (.24) was significantly greater than the coefficient depicting the feedback effect of behavioral engagement on changes in emotional engagement (.10,  $t = 2.90$ ,  $p < .01$ ). Contrary to predictions, however, for disaffection the feedforward effects of emotion on behavior were not stronger than the feedback effects of behavior on emotion.

#### Potential Facilitators of Engagement: SSPs

The third set of analyses focused on self-perceptions as potential facilitators of engagement. Concurrent correlations among the four indicators of engagement and three SSPs within the two time points appear in Table 4. As can be seen, all were in the predicted direction and significant at  $p < .001$ . To examine predictors of

Table 3  
 Correlations Among the Indicators of Engagement

Construct	1	2	3	4	5	6	7
1. Behavioral engagement	—	-.50	.60	-.35	-.43	-.23	-.21
2. Behavioral disaffection	-.44	—	-.42	.56	.52	.55	.42
3. Emotional engagement	.57	-.40	—	-.45	-.50	-.29	-.31
4. Emotional disaffection	-.36	.55	-.53	—	(.72)	(.87)	(.85)
5. Bored	-.44	.52	-.54	(.73)	—	.39	.46
6. Anxious	-.25	.47	-.36	(.86)	.40	—	.68
7. Frustrated	-.21	.39	-.38	(.84)	.46	.64	—

Note.  $N = 805$ . Correlations between variables within the Fall time point are below the diagonal. Correlations between variables within the Spring time point are above the diagonal. Correlations in parentheses are between subscale and total scale scores. All correlations are significant at  $p < .001$ .

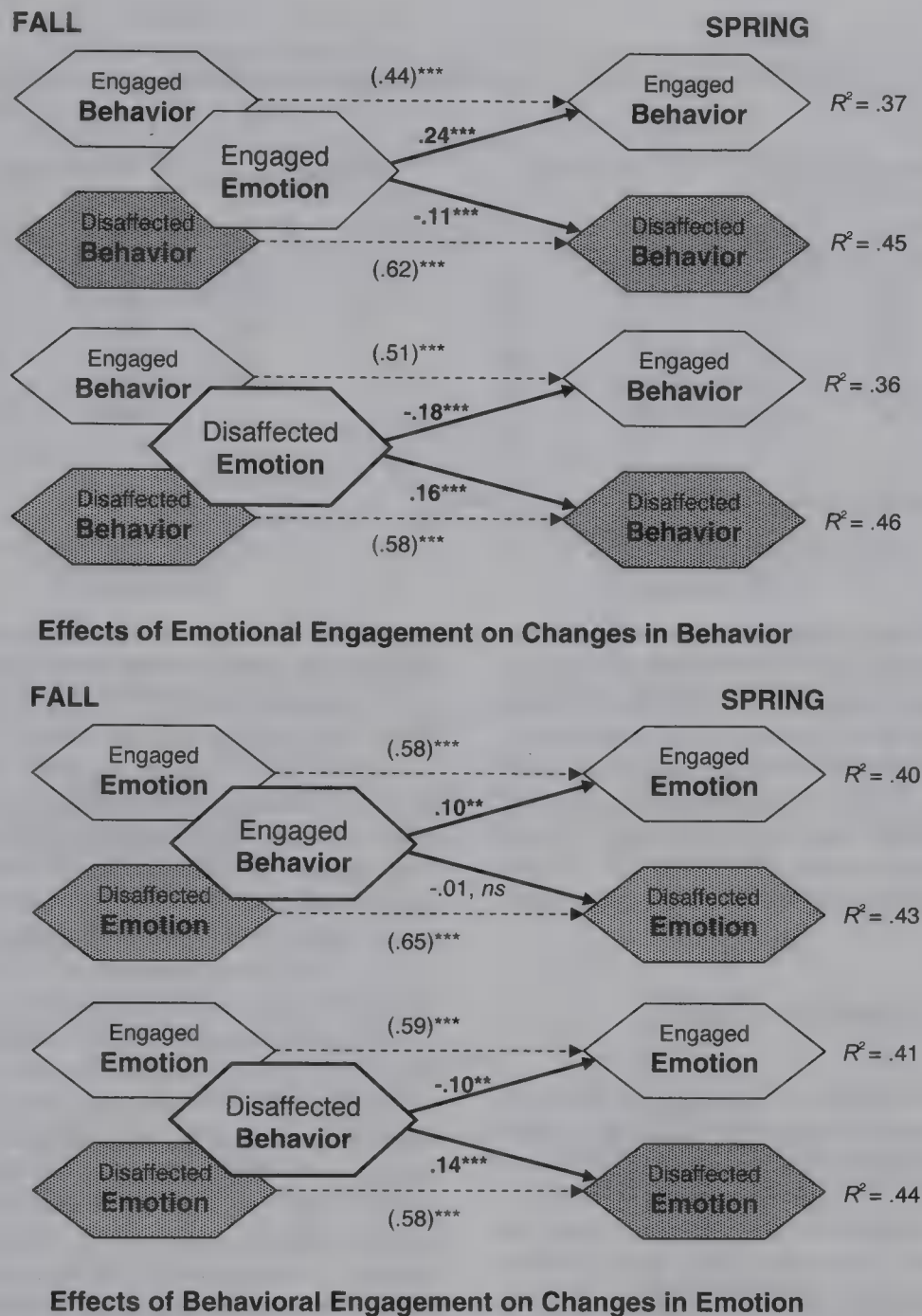


Figure 4. Results of regressions examining the internal dynamics of engagement. Top panel depicts the effects of emotional engagement and disaffection in the fall on changes in behavioral engagement and disaffection from fall to spring. Bottom panel depicts the effects of behavioral engagement and disaffection in the fall on changes in emotional engagement and disaffection from fall to spring.

Note. Standardized regression coefficients are on the solid arrows. Stability correlations from fall to spring are in parentheses.  $**p < .01$ .  $***p < .001$ .

change over the school year, we calculated multiple regressions in which we controlled for the dependent variable in the fall before examining whether the independent variable in fall predicted the dependent variable in spring.

*Self-systems as predictors of behavioral engagement and disaffection.* Figure 5 depicts the results of regressions examining whether SSPs contribute to changes in engagement and disaffection. As can be seen in the top panel, which presents their effects on the behavioral indicators, despite the high stabilities of the dependent variables, each self-system in the fall was a significant predictor of improvements in behavioral engagement and declines

in behavioral disaffection over the school year. Autonomy appeared to be the strongest predictor of change.

*Self-systems as predictors of emotional engagement and disaffection.* As shown in the bottom panel of Figure 5, despite the high stabilities of the emotional indicators, each self-system in the fall was a significant predictor of increases in emotional engagement over the year; again, autonomy was the strongest predictor. For emotional disaffection, however, autonomy was the only significant predictor of declines; despite robust concurrent correlations, neither relatedness nor perceived control were significant predictors of decreases in this indicator. Exploratory analyses ex-



Table 4  
Correlations Between Indicators of Engagement, Self-System Processes, and Teacher Support

Construct	Teacher support									
	Perceived control		Autonomy		Relatedness		Student report		Teacher report	
	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring
Behavioral engagement	.46	.53	.44	.42	.37	.37	.48	.54	.18	.21
Behavioral disaffection	-.49	-.54	-.47	-.45	-.33	-.36	-.44	-.50	-.24	-.22
Emotional engagement	.42	.41	.55	.53	.58	.51	.61	.60	.14	.12
Emotional disaffection	-.39	-.45	-.56	-.53	-.46	-.40	-.52	-.51	-.15	-.10
Bored	-.34	-.32	-.54	-.59	-.45	-.41	-.57	-.55	-.12	-.08*
Anxious	-.39	-.42	-.36	-.37	-.33	-.31	-.36	-.36	-.13	-.11
Frustrated	-.25	-.32	-.38	-.47	-.35	-.37	-.39	-.36	-.15	-.05

Note.  $N = 805$ . All correlations are significant at  $p < .001$ , except as indicated.

\*  $p < .05$ .

amined whether certain SSPs were stronger predictors of changes in the different disaffected emotions. We were particularly interested in whether children low in autonomy were at risk for increased boredom, whereas students with low perceived control might grow more anxious as the year progressed. Both relatedness ( $\beta = -0.08$ ,  $p < .05$ ) and autonomy ( $\beta = -0.12$ ,  $p < .001$ ) were significant predictors of decreases in boredom. However, perceived control was the only significant predictor of changes in anxiety ( $\beta = -0.08$ ,  $p < .05$ ), and autonomy was the only significant predictor of changes in frustration ( $\beta = -0.15$ ,  $p < .001$ ).

#### Potential Facilitators of Engagement: Teacher Supportive Context

The fourth set of analyses focused on the social facilitators of engagement. Both student reports of their interactions with teachers and teacher reports of their interactions with individual students were used as markers of the support provided by teachers. Correlations between the two markers of teacher support and the four indicators of engagement at two time points appear in Table 4. As can be seen, all were in the predicted direction and were generally significant at  $p < .001$ . Because engagement was also reported by students, student reports of teacher support were much more highly correlated with all indicators of engagement (average  $r = 1.531$ ) than were teacher reports of teacher support (average  $r = 1.171$ ). It is also possible that student reports reflect the more powerful causal influence because it is likely that students' perceptions of their interactions with teachers are the proximal causes of their motivational reactions. It should be noted that student and teacher reports of teacher support were positively correlated at both time points ( $r = .23$  in fall and  $r = .22$  in spring), indicating that despite differences in perspective and developmental level, the two reporters' assessments converged somewhat. Although few studies have examined cross-reporter consistency in perceptions of teacher support, the correlations found in the present study are within the typical range (e.g., Skinner & Belmont, 1993).

*Teacher support as a predictor of changes in engagement.* To examine whether teacher support contributed to changes in student engagement, we calculated two sets of multiple regressions in which either teacher or student reports of teacher support in fall were used to predict each of the indicators of engagement in spring, controlling for

that indicator in fall. Results for student reports of teacher support are depicted in Figure 6: In the top panel are findings for the regressions in which changes in behavioral indicators were the criterion; in the bottom panel are the results for changes in the emotional indicators. As expected, student reports of teacher support predicted improvements in emotional and behavioral engagement and declines in behavioral and emotional disaffection over time. In regressions examining whether teacher support had differential effects on the disaggregated disaffected emotions, we found that student reports of teacher support in fall predicted declines in both boredom ( $\beta = -0.14$ ,  $p < .001$ ) and frustration ( $\beta = -0.09$ ,  $p < .001$ ) from fall to spring; it did not predict changes in anxiety.

The findings for teacher reports of teacher support, although contrary to expectations, were not surprising, considering the strong cross-time stabilities of indicators of engagement and the modest concurrent correlations between teacher reports of teacher support and student-reported engagement. Changes in only one indicator of engagement were predicted by teacher reports of teacher support, namely, behavioral engagement ( $\beta = 0.07$ ,  $p < .05$ ). Consistent with findings for student reports, teacher reports indicated that students who received more teacher support in the beginning of the school year were likely to show improvements in their effort, attention, and persistence in the classroom as the year progressed.

#### Process Models of Potential Facilitators of Engagement

The final sets of analyses examined process models of the facilitators of engagement, in which actual teacher support (captured by teacher reports of teacher support) predicted changes in engagement by shaping students' perceptions of teacher support (captured by student reports of teacher support), which in turn contributed to children's feelings of competence, autonomy, and relatedness, which were themselves the proximal predictors of engagement. A mediated model posits that there would be no significant direct paths from either marker of teacher support to any indicator of engagement. We examined the two parts of the mediator models separately, using the four-step procedure recommended by Baron and Kenny (1986).

*Student experiences of teacher support as mediators of the effects of actual teacher support on changes in engagement.*

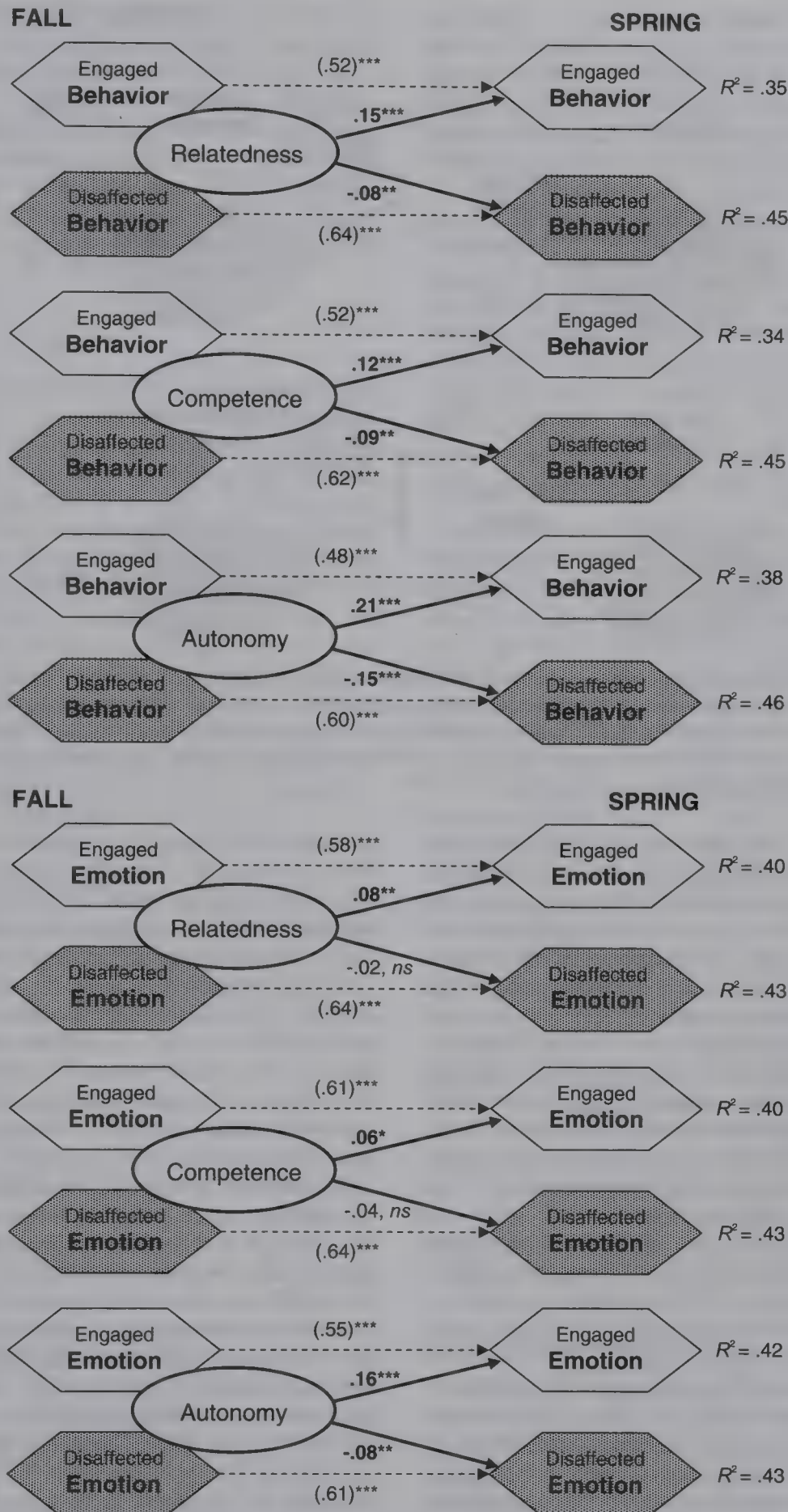


Figure 5. Results of regressions examining the effects of relatedness, competence, and autonomy in the fall on changes in engagement and disaffection from fall to spring. The top panel depicts changes in behavioral engagement and disaffection. The bottom panel depicts changes in emotional engagement and disaffection. Note. Standardized regression coefficients are on the solid arrows. Stability correlations from fall to spring are in parentheses. \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .



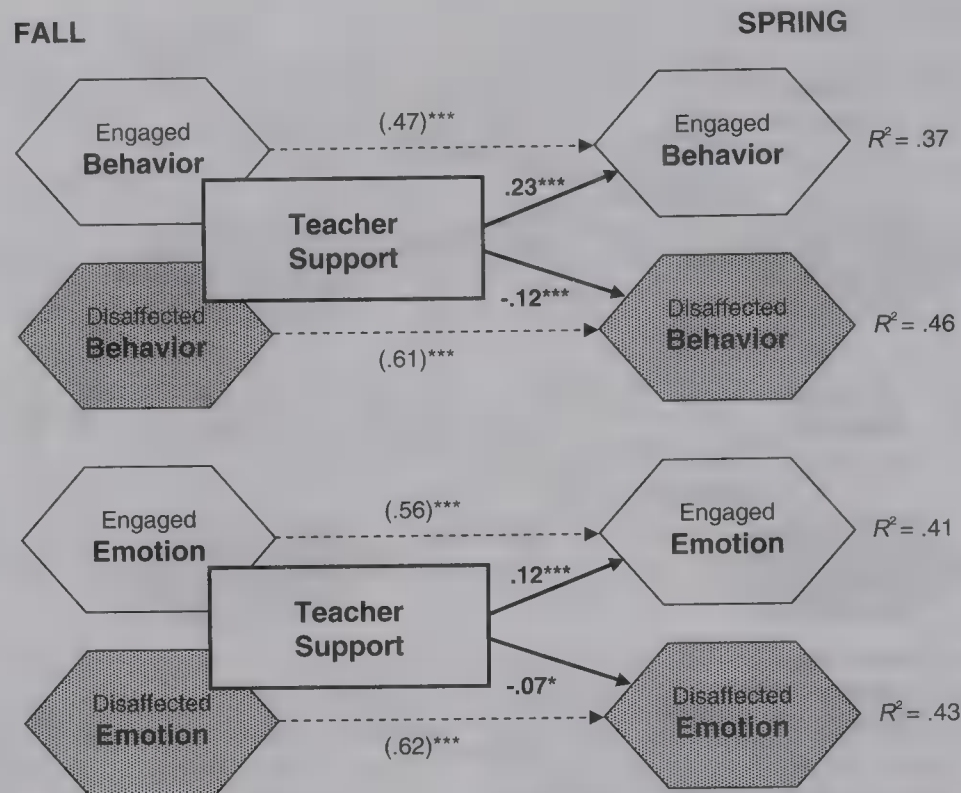


Figure 6. Results of regressions examining the effects of teacher support (student report) in the fall on changes in engagement and disaffection from fall to spring. The top panel depicts changes in behavioral engagement and disaffection. The bottom panel depicts changes in emotional engagement and disaffection. Note. Standardized regression coefficients are on the solid arrows. Stability correlations from fall to spring are in parentheses. \*  $p < .05$ . \*\*\*  $p < .001$ .

First, we explored whether the effects of actual teacher support (marked by teacher reports) on changes in behavioral engagement were mediated by students' perceptions of their interactions with teachers. Preliminary conditions for testing this model were met, namely, that the antecedent (i.e., teacher report of teacher support) was correlated with both (a) the outcome (i.e., changes in behavioral engagement) and (b) the proposed mediator (i.e., student report of teacher support) and that the proposed mediator was correlated with the outcome. The step of most interest was whether, in a regression using both the antecedent and the mediator as independent variables to predict the outcome, the unique effect of the mediator would remain significant, whereas the unique effect of the antecedent would be significantly reduced (indicating partial mediation) or no longer reach significance (indicating full mediation).

The model revealed full mediation. The effects of teacher report of teacher support on changes in behavioral engagement ( $\beta = 0.07, p < .05$ ) dropped substantially when student perceptions of teacher support were added to the equation and were no longer significant ( $\beta = 0.04, ns$ ), whereas student report of teacher support continued to be a significant unique predictor of changes in behavioral engagement ( $\beta = 0.22, p < .001$ ). This pattern of findings is consistent with the idea that teacher support shapes changes in students' behavioral engagement through its effects on children's perceptions of the support teachers provide. Mediational models for the other three indicators of engagement could not be tested because teacher report of teacher support was not a significant predictor of changes in any of them.

*Children's SSPs as mediators of the effects of teacher support on changes in engagement.* Second, we explored whether the effects of students' perceptions of teacher support on changes in the

four indicators of engagement were themselves mediated by students' SSPs of relatedness, competence, and autonomy. For these analyses, we created an aggregate marker by averaging the three SSPs together and followed the same procedures as in the last set of mediational models. Each model focused on changes in one indicator of engagement. For behavioral engagement, the models revealed partial mediation: The effect of teacher support on changes in behavioral engagement ( $\beta = 0.23, p < .001$ ) dropped substantially when the self-system aggregate was added to the equation, but it still remained significant ( $\beta = 0.13, p < .05$ ). As expected, the combined SSPs continued to be a significant unique predictor of changes in behavioral engagement ( $\beta = 0.15, p < .01$ ).

For emotional engagement, the models revealed full mediation: The effect of teacher support on changes in emotional engagement ( $\beta = 0.12, p < .001$ ) dropped substantially when the self-system aggregate was added to the equation and was no longer significant ( $\beta = 0.03, ns$ ), whereas the combined SSPs continued to be a significant unique predictor of changes in emotional engagement ( $\beta = 0.14, p < .001$ ). Similarly, the models for behavioral disaffection also revealed full mediation: The effect of teacher support on changes in behavioral disaffection ( $\beta = -0.12, p < .001$ ) dropped substantially when the self-system aggregate was added to the equation and was no longer significant ( $\beta = -0.04, ns$ ), whereas the combined SSPs continued to be a significant unique predictor of changes in emotional engagement ( $\beta = -0.12, p < .001$ ).

For emotional disaffection, neither predictor remained significant when the other variable was entered in the model: The effects of teacher support on changes in emotional disaffection ( $\beta = -0.07, p < .001$ ) was no longer significant when the aggregate self-systems were added to the equation ( $\beta = -0.06, ns$ ), but

neither were the effects of the combined SSPs ( $\beta = -0.02$ , *ns*). Because each variable was a significant predictor of changes in emotional disaffection when considered alone, the lack of significant unique effects can be attributed to multicollinearity. Taken together, these models suggest that teacher support has an effect on student engagement by shaping students' feelings of relatedness, competence, and autonomy. At the same time, teacher support (at least as perceived by students) also shows a direct effect on behavioral engagement, over and above that of the SSPs assessed in this study.

## Discussion

Guided by a motivational model, this study organized data on indicators and facilitators of engagement collected at the beginning and end of the school year to begin answering questions about the motivational dynamics of engagement. Consistent with other research on engagement (Fredricks et al., 2004; Wigfield et al., 2006), the developmental pattern found in this study, consisting of cross-year declines and age differences favoring younger children, reveal that although these students were relatively engaged, they nevertheless were experiencing losses in engagement and increases in disaffection over the transition to middle school. At the same time, the high cross-year stabilities indicate that children were losing ground commensurate with their initial levels of motivation. Systematic analyses of change over time suggest that this pattern may be fueled by the internal dynamics of engagement and by the larger motivational dynamics of which engagement, along with self-systems and teacher support, is a part.

### *Emotional Dynamics of Engagement*

Consistent with models of self-determination (Deci & Ryan, 1985), effectance motivation (Harter, 1978), and burgeoning interest in the role of emotions in the classroom (e.g., Pekrun, Goetz, Titz, & Perry, 2002; Schutz & DeCuir, 2002), findings from this study revealed positive emotions as one possible driver of children's effortful involvement in learning activities. At the same time, emotional disaffection, especially boredom, seemed to exert a significant downward pressure on children's effort and persistence and predicted their withdrawal from academic tasks. This pattern of findings underscores the idea that when children find learning activities interesting, fun, and enjoyable, they will pay more attention and try harder. However, the time lag from the beginning to the end of the school year suggests that this is more than a short-term gain. Children who are more emotionally engaged in fall show increasing behavioral engagement and declines in behavioral disaffection over the course of the whole year. By the same token, when children have lost their emotional enjoyment and interest in learning, they are not able to sustain behavioral participation in academic activities over time.

### *Behavioral Dynamics of Engagement*

The role of behavior in reciprocally shaping emotion was not as pronounced. Higher levels of effortful involvement in fall were not strong predictors of increases in emotional engagement, and they were not effective in staving off overall emotional disaffection, although they did predict declines in boredom when considered

separately. Behavioral disaffection, which in this study was marked by avoidance and withdrawal of effort, did seem to undermine the development of positive emotions and feed into increases in boredom, anxiety, and frustration. It makes sense that children with low classroom participation will eventually lose their enjoyment of learning activities and become more bored, anxious, and frustrated as the year progresses, whereas children who try hard and persist in learning activities will tend to find them increasingly more fun and enjoyable. However, it is unfortunate that such sustained involvement does not seem to pay off in reductions in anxiety or frustration over the long haul. This suggests that educators' efforts to increase behavioral engagement in ways that do not engage positive emotions may not have the intended lasting effect on children's high-quality participation in learning activities.

### *Larger Motivational Dynamics*

Engagement itself was also shaped by children's self-systems and by the support provided by their teachers. The clearest contributor to engagement was a sense of autonomy. Autonomy was a particularly strong predictor of changes in emotional engagement and disaffection—especially, as expected, of changes in boredom and frustration. Children who started the school year high in autonomy were likely to show improvements in their effort and enjoyment as the year progressed, whereas children low in autonomy (who felt externally or internally pressured) were likely to show increasing disaffection, both withdrawing their behavioral participation and feeling increasingly more bored and frustrated.

Competence made the strongest contributions to behavioral engagement and disaffection. Although not as strong a contributor to changes in emotional engagement overall, competence did seem to be a primary predictor of anxiety. Children with initially high levels of efficacy were likely to show improvements in their effort and exertion in class and to express moderate increases in their enjoyment and interest in learning activities over the school year. In contrast, children who started the school year doubting their capacities evinced increasing behavioral withdrawal from classwork accompanied by escalations in anxiety. Relatedness played a more central role in both kinds of engagement, but was not as strong a protective factor against disaffection. Children who began the year secure in their relationships with teachers increased in their effort and enjoyment, whereas children with less secure relationships with teachers were somewhat more likely to withdraw their efforts and to express boredom as the year progressed.

### *Role of Teacher Support*

The findings of this study suggest that teacher support also plays a central role in the motivational dynamics of engagement. If teacher reports of the support they provide are seen as markers of actual teacher support, then findings showed a pattern in which actual support is more important to behavioral engagement and disaffection (than to its emotional counterparts), whereas children's perceptions play a role in shaping changes in all facets of their participation in the classroom. Moreover, mediational analyses suggested that the contributions of teachers' support to changes in students' behavioral engagement are transmitted through their effects on children's perceptions of their interactions



with teachers. Students' perceptions of teacher support seemed to contribute to changes in engagement over the school year by shaping children's views of themselves as competent, autonomous, and related to teachers.

### *Self-System Model of Motivational Development*

The results of this study are largely congruent with the motivational model that suggests that teacher support, through its effects on students' perceptions of their interactions with teachers, shapes student self-systems over the school year, which in turn are strong predictors of all facets of their engagement. Engagement itself takes on its own dynamics, in which engagement and disaffection, through the reciprocal effects of behavior and emotion, tend to amplify themselves over time. These dynamics may explain the high interindividual stability within the motivational system, accompanied by slow declines that are portioned out differentially depending on initial levels of motivational resources. It should be noted that these connections do not differ as a function of gender or grade. The motivational processes that underlie the correlational results, although played out at different mean levels, seem to characterize all the gender and age groups examined in this study.

### *Implications for Conceptualizations of Engagement*

This study speaks directly to a set of issues raised in current reviews of the construct of engagement: how the components of engagement are similar and different from each other and how they work together over time (Fredricks et al., 2004; Jimerson et al., 2003). On one hand, the four indicators of engagement are similar in many ways: They are all relatively stable over the school year, at the same time that they are all worsening somewhat. They are all shaped both by teacher support and by students' self-perceptions, especially of autonomy, and the four indicators are all correlated with each other highly enough that they could reasonably be combined to form a single internally consistent bipolar construct.

On the other hand, each component has its own distinctive antecedents and its own role in the internal dynamics of engagement. The core construct, most prototypical of engagement, is behavioral participation in the classroom. In this sample, it had the highest mean level, the lowest cross-time stability, and the fastest drop across the school year. Behavioral engagement registered the biggest internal effects from emotional engagement and disaffection and from each of the differentiated disaffected emotions, especially boredom. It also registered the biggest effect from the SSPs, especially competence, and was the only indicator shaped by both teacher and student reports of teacher support. Hence, behavioral engagement seems to be a good summary indicator, diagnostic of the state of the entire motivational system. At the same time, however, it is not a strong contributor to changes in the other facets of engagement—it is not a big booster of subsequent enjoyment and interest, nor can it forestall emotional disaffection.

Compared with behavioral engagement, emotional engagement has a somewhat different profile: It is lower in mean level, a bit more stable, and loses less ground over the school year. However, it also seems to be a sensitive barometer of the whole motivational system, as emotional engagement was shaped over time by each SSP (with especially strong contributions from relatedness and autonomy) and also by students' perceptions of their interactions with teachers. Most

important, emotional engagement appears to be the active ingredient in sustaining motivation: It is the strongest contributor to the feedforward internal dynamics of engagement, bolstering behavioral engagement and staving off behavioral disaffection.

The two kinds of disaffection, which were both relatively low and interindividually stable, nevertheless seemed to feed on each other over time. Children who felt emotionally disaffected withdrew their effort over the school year, and as children stop participating, they became more emotionally alienated. Teacher support and the SSPs also played a role, especially in shaping behavioral disaffection. Analyses of the differentiated disaffected emotions suggest that children low in autonomy and relatedness are especially at risk for developing boredom; children low in perceived control, for escalating anxiety; and students low in autonomy, for increasing frustration.

### *Limitations of the Present Study*

Before discussing the implications of the findings further, the limitations of the current study must be taken into account. In terms of participants, this study focused on a group of middle- and working-class students who were largely Caucasian and were drawn from only two schools. Of course, it is noteworthy that modest declines in engagement were found even for this well-functioning group of students. However, it is important to be cautious about generalizing these results to subgroups who show the steepest declines in motivation. Even though many of the same findings have been documented in African American and low-income groups (Connell et al., 1994, 1995; Gutman & Midgley, 2000), little is known about the internal dynamics of engagement for these (or any other) children.

In terms of measures, the study relied heavily on student self-reports. Of course, students may be the only source for information about their SSPs, but it would have been helpful to include observational assessments of engagement and of interactions with teachers. In terms of design, the use of two points of measurement was a decided improvement over a one-time assessment. Nevertheless, time points at the beginning and end of the school year have no particular correspondence to the kind of episodic time during which these motivational cycles are hypothesized to unfold. It is more likely, for example, that the effects of student-teacher interactions on children's subsequent engagement play out over a period of weeks or months. A design that incorporates more frequent time intervals would be better suited to capture these dynamics (e.g., Schmitz & Skinner, 1993).

### *Implications for Theoretical Development and Future Research*

This study contributes to our growing understanding of the indicators and facilitators of engagement. The distinctions proposed between behavioral and emotional engagement and disaffection allowed for the study of how these components shape each other over time and uncovered enough differences in their operation to suggest that in future studies researchers must carefully consider whether they can be meaningfully combined. At the same time, these markers of classroom participation could themselves be distinguished from the self-systems and qualities of student-teacher interactions that have sometimes been included under the



broad conceptual sweep of "engagement" (Jimerson et al., 2003). Even if social contexts and self-perceptions are relevant and important (an assertion supported by this study), they should be unpacked from engagement itself if research is to investigate how these personal and social factors have an impact on the quality of students' participation in learning activities over time.

This study also suggests that future empirical efforts can build on the SSMMD by incorporating additional factors drawn from reviews of motivation and engagement (e.g., Wigfield et al., 2006). Other facets of engagement, especially cognitive forms, can be considered in relation to behavioral and emotional features, along with more disruptive forms of disaffection (Fredricks et al., 2004). Moreover, the general motivational dynamics are also likely to include additional components, for example, other facets of teachers' actions (such as competence feedback and their own enthusiasm) and the nature of the learning tasks students are required to undertake (especially the extent to which they are interesting, fun, and relevant). Other self-perceptions (e.g., goal orientations or values) can be incorporated as predictors of engagement and perhaps as alternative pathways through which teachers shape student motivation (Brophy, 2004; Stipek, 2002).

Perhaps most important, future studies can incorporate the assessment of important long-term motivational outcomes, such as identification with and commitment to school (Deci & Ryan, 1985; Finn, 1989, 1993), the internalization of the values of achievement and learning (Brophy, 1999; Eccles & Wigfield, 2002), and the development of the capacities and motivation for self-regulated learning (Ryan & Connell, 1989; Schunk & Zimmerman, 1994) or coping with academic difficulties and challenges (Skinner & Wellborn, 1994, 1997). These are the enduring motivational resources to which students' active enthusiastic participation likely contributes during late elementary school and that may act as protective buffers as students go through the normatively challenging transitions to middle and high school.

In sum, the general direction of this research appears promising as a source of insights about the dynamics of student engagement. The study suggests, for example, the centrality of children's interest and emotion in initiating and sustaining their participation in learning activities (e.g., Pekrun et al., 2002; Schutz & DeCuir, 2002) and highlights the burden that an emotion as commonplace as boredom can put on children's effortful involvement in academic tasks. Our results underscore the importance of student autonomy to engagement (Deci et al., 1985) as well as pointing out that its low mean level suggests that this need is not well met, even in this generally well-functioning group. At the same time, all of the self-systems we considered play a role—feelings of competence are needed to bolster exertion and persistence, and relatedness and autonomy are needed to spark the interest and enjoyment that sustains effort over time. Taken together, our findings suggest that the behaviors and emotions students present in class may provide teachers with a window on the inner workings of children's motivational resources and vulnerabilities (Furrer, Kelly, & Skinner, 2003). Patterns of engagement and disaffection may be diagnostic of the state of students' feelings of relatedness, competence, and autonomy, and if they are faltering, teachers may be able to figure out the kinds of motivational supports that could bolster them. Such compensatory teacher reactions may suggest one avenue for helping diminish or perhaps even reverse self-amplifying cycles of disaffection. Future studies focusing on these

and other questions of motivational dynamics may further elucidate the role of engagement in the long-term development of student academic resilience and success.

## References

- Ainsworth, M. D. S., Blehar, E., Waters, E., & Wall, S. (1978). *Patterns of attachment*. Hillsdale, NJ: Erlbaum.
- Anderman, L. H. (1999). Classroom goal orientation, school belonging, and social goals as predictors of students' positive and negative affect following transition to middle school. *Journal of Research and Development in Education*, 32, 89–103.
- Anderman, L. H., & Anderman, E. M. (1999). Social predictors of changes in students' achievement goal orientations. *Contemporary Educational Psychology*, 24, 21–37.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.
- Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173–1182.
- Battistich, V., Solomon, D., Watson, M., & Schaps, E. (1997). Caring school communities. *Educational Psychologist*, 32, 137–151.
- Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117, 497–529.
- Birch, S. H., & Ladd, G. W. (1997). The teacher-child relationship and children's early school adjustment. *Journal of School Psychology*, 35, 61–79.
- Birch, S. H., & Ladd, G. W. (1998). Children's interpersonal behaviors and the teacher-child relationship. *Developmental Psychology*, 34, 934–946.
- Bowlby, J. (1969/1973). *Attachment and loss. Vols. 1 and 2*. New York: Basic Books.
- Brophy, J. E. (1999). Toward a model of the value aspects of motivation in education: Developing appreciation for particular learning domains and activities. *Educational Psychologist*, 34, 75–85.
- Brophy, J. E. (2004). *Motivating students to learn* (2nd ed.). Mahwah, NJ: Erlbaum.
- Connell, J. P., Halpern-Felsher, B. L., Clifford, E., Crichlow, W., & Usinger, P. (1995). Hanging in there: Behavioral, psychological, and contextual factors affecting whether African-American adolescents stay in high school. *Journal of Adolescent Research*, 10, 41–63.
- Connell, J. P., Spencer, M. B., & Aber, J. L. (1994). Educational risk and resilience in African-American youth: Context, self, action, and outcomes in school. *Child Development*, 65, 493–506.
- Connell, J. P., & Wellborn, J. G. (1991). Competence, autonomy and relatedness: A motivational analysis of self-system processes. In M. Gunnar & L. A. Sroufe (Eds.), *Minnesota Symposium on Child Psychology: Vol. 23. Self processes in development* (pp. 43–77). Chicago: University of Chicago Press.
- Deci, E. L., Connell, J. P., & Ryan, R. M. (1985). A motivational analysis of self-determination and self-regulation in the classroom. In C. Ames & R. Ames (Eds.), *Research on motivation in education: Vol. 2. The classroom milieu* (pp. 13–52). San Diego, CA: Academic Press.
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum Press.
- Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. Philadelphia: Psychology Press.
- Eccles, J. S., & Midgley, C. (1989). Stage-environment fit: Developmentally appropriate classrooms for early adolescents. In R. E. Ames & C. Ames (Eds.), *Research on motivation in education: Goals and cognitions* (Vol. 3, pp. 13–44). New York: Academic Press.
- Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, 53, 109–132.



- Elliot, A. J., & Dweck, C. S. (Eds.). (2005). *Handbook of competence and motivation*. New York: Guilford Press.
- Finn, J. D. (1989). Withdrawing from school. *Review of Educational Research*, 59, 117-142.
- Finn, J. D. (1993). *School engagement and students at risk* (NCES 93470). Washington, DC: National Center for Education Statistics.
- Finn, J. D., Pannozzo, G. M., & Voelkl, K. E. (1995). Disruptive and inattentive-withdrawn behavior and achievement among fourth graders. *Elementary School Journal*, 95, 421-454.
- Finn, J. D., & Rock, D. A. (1997). Academic success among students at risk for school failure. *Journal of Applied Psychology*, 82, 221-234.
- Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 74, 59-109.
- Furrer, C., Kelly, G., & Skinner, E. (2003, April). *Can teachers use children's emotions in the classroom to diagnose and treat underlying motivational problems?* Poster presented at the biennial meetings of the Society for Research in Child Development, Tampa, FL.
- Furrer, C., & Skinner, E. A. (2003). Sense of relatedness as a factor in children's academic engagement and performance. *Journal of Educational Psychology*, 95, 148-162.
- Furrer, C., Skinner, E., Marchand, G., & Kindermann, T. A. (2006, March). *Engagement versus disaffection as central constructs in the dynamics of motivational development*. Paper presented at the Society for Research on Adolescence, San Francisco, CA.
- Goodenow, C. (1993). Classroom belonging among early adolescent students: Relationships to motivation and achievement. *Journal of Early Adolescence*, 13, 21-43.
- Grolnick, W. S., & Ryan, R. M. (1987). Autonomy in children's learning: An experimental and individual difference investigation. *Journal of Personality and Social Psychology*, 52, 890-898.
- Guthrie, J. T., & Davis, M. H. (2003). Motivating struggling readers in middle school through an engagement model of classroom practice. *Reading & Writing Quarterly*, 19, 59-85.
- Gutman, L., & Midgley, C. (2000). The role of protective factors in supporting the academic achievement of poor African American students during the middle school transition. *Journal of Youth and Adolescence*, 29, 223-248.
- Hardre, P. L., & Reeve, J. (2003). A motivational model of students' intentions to persist in, versus drop out of, high school. *Journal of Educational Psychology*, 95, 347-356.
- Harter, S. (1978). Effectance motivation reconsidered: Toward a developmental model. *Human Development*, 21, 36-64.
- Harter, S. (1982). The perceived competence scale for children. *Child Development*, 53, 89-97.
- Jimerson, S. R., Campos, E., & Greif, J. L. (2003). Toward an understanding of definitions and measures of school engagement and related terms. *California School Psychologist*, 8, 7-27.
- Kuperminc, G. P., Blatt, S. J., Shahar, G., Henrich, C., & Leadbetter, B. J. (2004). Cultural equivalence and cultural variance in longitudinal associations of young adolescent self-definition and interpersonal relatedness to psychological and school adjustment. *Journal of Youth and Adolescence*, 33, 13-30.
- Lynch, M., & Cicchetti, D. (1997). Children's relationships with adults and peers: An examination of elementary and junior high school students. *Journal of School Psychology*, 35, 81-99.
- Meyer, D. K., & Turner, J. C. (2002). Discovering emotion in classroom research. *Educational Psychologist*, 37, 107-114.
- Miserandino, M. (1996). Children who do well in school: Individual differences in perceived competence and autonomy in above-average children. *Journal of Educational Psychology*, 88, 203-214.
- Murdock, T. B. (1999). The social context of risk: Status and motivational predictors of alienation in middle school. *Journal of Educational Psychology*, 91, 62-75.
- Murray, C., & Greenberg, M. T. (2000). Children's relationship with teachers and bonds with school. *Journal of School Psychology*, 38, 423-445.
- O'Farrell, S. L., & Morrison, G. M. (2003). A factor analysis exploring school bonding and related constructs among upper elementary students. *California School Psychologist*, 8, 53-72.
- Patrick, B. C., Skinner, E. A., & Connell, J. P. (1993). What motivates children's behavior and emotion? Joint effects of perceived control and autonomy in the academic domain. *Journal of Personality and Social Psychology*, 65, 781-791.
- Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, 37, 91-105.
- Pierson, L. H., & Connell, J. P. (1992). Effect of grade retention on self-system processes, school engagement, and academic performance. *Journal of Educational Psychology*, 84, 300-307.
- Reeve, J., Bolt, E., & Cai, Y. (1999). Autonomy supportive teachers: How they teach and motivate students. *Journal of Educational Psychology*, 91, 537-548.
- Reeve, J., Jang, H., Carrell, D., Jeon, S., & Barch, J. (2004). Enhancing students' engagement by increasing teachers' autonomy support. *Motivation and Emotion*, 28, 147-169.
- Roeser, R. W., Midgley, C., & Urdan, T. C. (1996). Perceptions of the school psychological environment and early adolescents' psychological and behavioral functioning in school: The mediating role of goals and belonging. *Journal of Educational Psychology*, 88, 408-422.
- Roeser, R., Strobel, K. R., & Quihuis, G. (2002). Studying early adolescents' academic motivation, social-emotional functioning, and engagement in learning: Variable- and person-centered approaches. *Anxiety, Stress, and Coping*, 15, 345-368.
- Ryan, A. M. (2000). Peer groups as a context for the socialization of adolescents' motivation, engagement, and achievement in school. *Educational Psychologist*, 35, 101-111.
- Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology*, 57, 749-761.
- Ryan, R. M., & Powelson, C. L. (1991). Autonomy and relatedness as fundamental to motivation and education. *Journal of Experimental Education*, 60(1), 49-66.
- Ryan, R. M., Stiller, J. D., & Lynch, J. H. (1994). Representations and relationships to teachers, parents, and friends as predictors of academic motivation and self-esteem. *Journal of Early Adolescence*, 14, 226-249.
- Schmitz, B., & Skinner, E. (1993). Perceived control, effort, and academic performance: Interindividual, intraindividual, and multivariate time-series analyses. *Journal of Personality and Social Psychology*, 64, 1010-1028.
- Schunk, D. H., & Zimmerman, B. J. (Eds.). (1994). *Self-regulation of learning and performance*. Hillsdale, NJ: Erlbaum.
- Schutz, P. A., & DeCuir, J. T. (2002). Inquiry on emotions in education. *Educational Psychologist*, 37, 125-134.
- Sinclair, M. F., Christenson, S. L., Lehr, C. A., & Anderson, A. R. (2003). Facilitating school engagement: Lessons learned from Check & Connect longitudinal studies. *California School Psychologist*, 8, 29-41.
- Skinner, E. A. (1995). *Perceived control, motivation, and coping*. Newbury Park, CA: Sage.
- Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology*, 85, 571-581.
- Skinner, E. A., Chapman, M., & Baltes, P. B. (1983). *The Control, Agency and Means-ends Interview (CAMI)*. (English and German versions, Technical Report). Berlin, Germany: Max Planck Institute.
- Skinner, E. A., Chapman, M., & Baltes, P. B. (1988). Control, means-ends, and agency beliefs: A new conceptualization and its measurement during childhood. *Journal of Personality and Social Psychology*, 54, 117-133.

- Skinner, E. A., Kindermann, T. A., Connell, J. P., & Wellborn, J. G. (in press). Engagement as an organizational construct in the dynamics of motivational development. In K. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school*. Mahwah, NJ: Erlbaum.
- Skinner, E. A., Kindermann, T. A., & Furrer, C. (in press). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and Psychological Measurement*.
- Skinner, E. A., & Wellborn, J. G. (1994). Coping during childhood and adolescence: A motivational perspective. In D. Featherman, R. Lerner, & M. Perlmutter (Eds.), *Life-span development and behavior* (Vol. 12, pp. 91–133). Hillsdale, NJ: Erlbaum.
- Skinner, E. A., & Wellborn, J. G. (1997). Children's coping in the academic domain. In S. A. Wolchik & I. N. Sandler (Eds.), *Handbook of children's coping with common stressors: Linking theory and intervention* (pp. 387–422). New York: Plenum Press.
- Skinner, E. A., Wellborn, J. G., & Connell, J. P. (1990). What it takes to do well in school and whether I've got it: The role of perceived control in children's engagement and school achievement. *Journal of Educational Psychology*, 82, 22–32.
- Skinner, E. A., Zimmer-Gembeck, M. J., & Connell, J. P. (1998). Individual differences and the development of perceived control. *Monographs of the Society for Research in Child Development*, 63(2–3), v–220.
- Stipek, D. J. (2002). *Motivation to learn: From theory to practice* (4th ed.). Needham Heights, MA: Allyn & Bacon.
- Vallerand, R. J., Fortier, M. S., & Guay, F. (1997). Self-determination and persistence in a real-life setting: Toward a motivational model of high school dropout. *Journal of Personality and Social Psychology*, 72, 1161–1176.
- Weiner, B. (1990). History of motivation research in education. *Journal of Educational Psychology*, 82, 616–622.
- Weiner, B. (2005). Motivation from an attributional perspective and the social psychology of perceived competence. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 73–84). New York: Guilford.
- Wellborn, J. G. (1991). *Engaged and disaffected action: The conceptualization and measurement of motivation in the academic domain*. Unpublished doctoral dissertation, University of Rochester.
- Wentzel, K. R. (1993). Does being good make the grade? Social behavior and academic competence in middle school. *Journal of Educational Psychology*, 85, 357–364.
- Wentzel, K. R. (1997). Student motivation in middle school: The role of perceived pedagogical caring. *Journal of Educational Psychology*, 89, 411–419.
- Wentzel, K. R. (1998). Social relationships and motivation in middle school: The role of parents, teachers, and peers. *Journal of Educational Psychology*, 90, 202–209.
- Wentzel, K. R. (1999). Social-motivational processes and interpersonal relationships: Implications for understanding motivation at school. *Journal of Educational Psychology*, 91, 76–97.
- White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review*, 66, 297–333.
- Wigfield, A., Eccles, J. S., Schiefele, U., Roeser, R., & Davis-Kean, P. (2006). Development of achievement motivation. In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (6th ed., pp. 933–1002). New York: Wiley.

## Appendix

### Engagement Versus Disaffection With Learning: Student-Report

#### Behavioral Engagement

1. I try hard to do well in school.
2. In class, I work as hard as I can.
3. When I'm in class, I participate in class discussions.
4. I pay attention in class.
5. When I'm in class, I listen very carefully.

#### Emotional Engagement

1. When I'm in class, I feel good.
2. When we work on something in class, I feel interested.
3. Class is fun.
4. I enjoy learning new things in class.
5. When we work on something in class, I get involved.

#### Behavioral Disaffection

1. When I'm in class, I just act like I'm working. (–)
2. I don't try very hard at school. (–)
3. In class, I do just enough to get by. (–)
4. When I'm in class, I think about other things. (–)
5. When I'm in class, my mind wanders. (–)

#### Emotional Disaffection

1. a. When we work on something in class, I feel bored. (–)
- b. When I'm doing work in class, I feel bored. (–)

- c. When my teacher first explains new material, I feel bored. (–)
2. a. When I'm in class, I feel worried. (–)
- b. When we start something new in class, I feel nervous. (–)
- c. When I get stuck on a problem, I feel worried. (–)
3. When we work on something in class, I feel discouraged. (–)
4. Class is not all that fun for me. (–)
5. a. When I'm in class, I feel bad. (–)
- b. When I'm working on my classwork, I feel mad. (–)
- c. When I get stuck on a problem, it really bothers me. (–)
- d. When I can't answer a question, I feel frustrated. (–)

*Note.* Adapted from *Engaged and Disaffected Action: The Conceptualization and Measurement of Motivation in the Academic Domain*, by J. G. Wellborn, 1991, unpublished doctoral dissertation, University of Rochester. Copyright 1991 by J. G. Wellborn. Adapted with permission. The items added to the Emotional Disaffection subscale can be used to tap the more differentiated disaffected emotions.

Received March 20, 2007

Revision received April 21, 2008

Accepted April 22, 2008 ■



# Children's Early Interest-Based Activities in the Home and Subsequent Information Contributions and Pursuits in Kindergarten

Carin Neitzel  
Vanderbilt University

Joyce M. Alexander  
Indiana University at Bloomington

Kathy E. Johnson  
Indiana University–Purdue University at Indianapolis

This study examined the early interests of 109 children and their subsequent information contributions and pursuits in kindergarten. Four groups of children with similar interests were identified on the basis of the children's profiles of activities in the home, tracked bimonthly for over a year. Activity patterns reflected conceptual, social, procedural, or creative interests. The role of early interests in understanding academic engagement was investigated, with gender, cognitive skill, and temperament statistically controlled. Observational data from throughout the school year revealed differences in the types of information that children contributed to discussions and pursued in class related to children's early interests. Findings enrich understanding of young children's academic behaviors and extend theoretical models of academic self-instruction behaviors such as information exchanges and pursuits in classrooms.

**Keywords:** early childhood, personal interest, academic participation, information pursuits

Academically successful children institute an anthology of skills supportive of learning. Included among this repertoire of behaviors are self-initiated interactions with and quests for information (Zimmerman, 1989). For example, a good student actively contributes to scholarly discussions in the classroom and independently pursues information about academic topics, tasks, strategies, or performances, among other participation options. These interactions are a forum for students to access, elaborate, organize, integrate, or verify information and enhance understanding (Butler & Winne, 1995; Woloshyn, Pressley, & Schneider, 1992). The ability or willingness to adapt to the cognitive behavioral expectations of school has long-term as well as immediate effects on learning (Krapp & Fink, 1992). Consequently, it is important to understand the factors that may influence young children's academic uses and pursuits of information in the classroom.

---

Carin Neitzel, Department of Teaching and Learning, Vanderbilt University, Nashville, Tennessee; Joyce M. Alexander, Department of Educational Psychology, Indiana University at Bloomington, Bloomington, Indiana; Kathy E. Johnson, Department of Psychology, Indiana University–Purdue University at Indianapolis, Indianapolis, Indiana.

This research was supported by National Science Foundation Grant BCS-9907865 to Joyce M. Alexander and Kathy E. Johnson and by a Proffitt research endowment grant from the School of Education at Indiana University. Thank you to Fabiola Reis-Henrie, Mary E. Leibham, Che-yu Kuo, Khidhir Zakaria, Christy Kendrick, and other research assistants at IUPUI and IU for their assistance on this project. We are very grateful to the children and parents who have graciously supported this research through their participation. We also thank the principals and teachers who welcomed us into their classrooms.

Correspondence concerning this article should be addressed to Carin Neitzel, Department of Teaching and Learning, Vanderbilt University, GPC Box 230, 230 Appleton Place, Nashville, TN 37203. E-mail: carin.neitzel@vanderbilt.edu

Children's personalities, cognitive abilities, and motivational dispositions have all been linked to their academic behaviors (Bjorklund & Schneider, 1996), and interest has been credited as the bridge linking cognition, motivation, and academic outcomes (Hidi & Harackiewicz, 2000; Pekrun, Goetz, Titz, & Perry, 2002). Although numerous studies have investigated the ways in which psychological state variables related to learning such as focus, effort, and persistence are affected by individual interest, the impact of interest on young students' means of academic participation in school—such as self-initiated information contributions and pursuits—has been relatively unexamined.

Children as young as 3 years already express intense interests in particular forms of activity (Renninger, 1989). Interest occurs when a certain object or event triggers a curiosity response or sense of enjoyment (Fink, 1994). Some interest theorists have proclaimed that it is best defined by the content of activity rather than the object of activity because even if interest may be piqued by particular objects, different materials elicit and support different types of experiences and interactions or action potentials (Renninger, 1984). In turn, interest influences not only the way a child engages in a given activity but also the child's representation of interaction possibilities (Renninger, 1990). Consequently, early interest-driven activity types may render certain transactions more familiar or some forms of information more valuable, and future interactions may be informed by early kinds of engagement (Renninger, 1990, 2000).

The study of interest from an action-potential perspective has received relatively little attention in empirical research. To date, the type of interest most heavily investigated has been topic interest (Schiefele, 2001), and in queue, studies of interest outcomes have focused almost exclusively on its influence on children's efficiency or effectiveness in learning subject- or domain-relevant information (Hidi & Anderson, 1992; Hidi, Berndorff, &

Ainley, 2002; Schiefele, 1990, 2001) or intensity and duration of involvement, dispositions, and motivation to learn in areas of interest (Hidi, 1990; Hidi & McLaren, 1991; Renninger & Leckrone, 1991; Schiefele, 2001). Studies of preschoolers have noted their use of interest objects as adaptive tools or security anchors and use of interest topics to initiate social contact with other children as they transition from home to school (Baum, Renzulli, & Hebert 1995; Krapp & Fink, 1992). However, the influence of interest may transcend content domains and exert more pervasive influence on children's participation patterns across settings. Early interests may direct subsequent participation strategy selection and use. Previous research with singular interest type foci has generated little information about how various interests may prepare children differently for interactions with academic information.

The present study examined children's interest-based activities in the home during early childhood and their subsequent contributions to academic discussions and quests for information in kindergarten. Temperament and cognitive characteristics of the child that could compete to explain differences in these academic behaviors were controlled. Consequently, this study offers an examination of the unique role of various early childhood interests for understanding differences in young children's uses and solicitations of information across domains and learning contexts in school.

### Children's Interests Expressed Through Play

Renninger and Wozniak (1985) suggested that children's interests can be identified by examining their time spent and level of engagement in particular activities relative to others. The activities in which young children consistently involve themselves are believed to reflect general orientations or socialized preferences for particular features of engagement (Krapp, Hidi, & Renninger, 1992; Renninger, 1989). Recent research has supported the notion that young children's pattern of activities may be indicative of some underlying comprehensive interest (Renninger, 1989) and shows that children themselves categorize their play on the basis of features or attributes of particular activities (Holmes, 1991). In fact, consistency has been noted in the action types of young children even in engagements with different objects (Renninger & Leckrone, 1991; Saracho, 1995). Therefore, young children's profiles of activity may be a better indicator of their interests than time spent within any single area of play.

Previous research has identified three attributes of activity that characterize interests in early childhood: material versus social focus, property or process referent, and reproductive versus transformative potential (Renninger, 1984). Some children are more focused on material-based activities best described as property or concept oriented (Renninger, 1989). During concept-oriented activity, the child focuses on object domain-specific or topical information exploration (Fink, 1994). Other children are focused more on particular practices or actions with objects (Renninger, 1989), referred to as process-oriented activity. These activities vary in how prescribed the schemas or scripts are that guide them (Johnson, Christie, & Yawkey, 1987; Pulaski, 1970; Renninger, 1984). Some materials have prescribed uses and functions, and the schemas or scripts that guide such activity are limited. Other materials are relatively free from intrinsic limits or structure, and thus, schema or script variability and transformative potential are

high. Therefore, process-oriented activity may be either highly structured (*procedural*), or open ended (*creative*). In contrast to children with these material-based interest types, some children are focused more on activities best described as socially oriented (Jennings, 1975; Krapp & Fink, 1992; Renninger, 1984). During socially oriented activity, the child is focused on enactment of social roles, routines, or processes.

Topical interests have been most heavily studied (Hidi et al., 2002; Schiefele, 1990, 2001); socially situated interests also have received attention in recent research (Sansone & Smith, 2000). These categories of interest as well as procedural and creative interests have been utilized in previous studies of vocational interest (Gati, 1991; Tracey, 1997). Less frequently, these categories have been used to describe differences in young children's personal interest types (Renninger, 1990), and, to date, this research has primarily examined intraindividual differences in children's play preferences. Presuming that individuals do have activity attribute penchants—material or social focused, property or process based, and structured or transformative preferences—children with common interests likely exhibit systematic similarities in the types of activities in which they spend their play time. This study examined children's profiles of play behaviors in the home to investigate the effects of differences in children's early childhood play interests and their subsequent academic involvement behaviors (contributions to discussion and pursuits of information) in school.

### Early Interests and Children's Academic Behaviors

Most studies of student learning approaches have been guided by views of academic involvement as a set of cognitive skills or a self-system trait (Vermunt, 1996, 1998). Students' levels and forms of participation in school have been correlated with—though distinctive from—general measures of cognitive skill (Zimmerman & Bandura, 1994; Zimmerman & Martinez-Pons, 1986). Personality characteristics have been associated with students' willingness to make use of opportunities to engage and sources of information in school (Birch & Ladd, 1998). In addition, research has shown that boys tend to control classroom conversations more than girls in the early elementary school years (Brophy, 1985; Jovanovic & King, 1998). Although the presence or absence of interest has been associated with children's level of engagement in school tasks (Schiefele, 2001), differences in children's personal interests have not been linked to their use of particular involvement strategies in classrooms. However, different interests lead to experiences that support different types of interacting and thinking (Sigel, 1993) and impact what and how a child learns (Baum et al., 1995; Renninger, 1990, 1992; Sansone, Weir, Harpster, & Morgan, 1992) by influencing the types of information that the individual attends to, deeply processes, and encodes (Hidi, 1990; Renninger, 1990; Renninger & Wozniak, 1985) as well as the individual's perceptions of values and expectations for behavior (Tobias, 1994). In these ways, interest serves as a mental and affective resource and leads to an enriched knowledge base and increased motivation to engage in particular ways (Ericsson & Smith, 1991; Hidi, 1990) that may have implications for children's involvement strategy selection in new events and settings.



Contemporary theories of interest distinguish between *situational interests*, which are triggered by external stimuli, and *individualized interests*, which are characterized by a relatively enduring predisposition to interact within a target domain (Krapp, 2002; Renninger & Hidi, 2002). However, theorists also assert that situational factors can both "catch" interest initially and "hold" it over time (Krapp, 2002) and that even individual interests are socially constructed and therefore are situational to some degree (Pressick-Kilborn & Walker, 2002). Whether intrinsically or extrinsically piqued, early interest-based activities maintained over time may exert some influence on subsequent interactions (Krapp, 2002). Therefore, differences in children's interest-based activities in early childhood may be related to the differences observed in children's contributions to academic discussions and information pursuits in kindergarten classrooms.

### *Contributions to Academic Discussions*

One's prior knowledge has a strong influence on cognitive engagement and forms of engagement with tasks (Winne, 1995). A well-developed knowledge base allows children to contribute to academic discussions in the classroom by equipping them with the tools with which they can work (Woloshyn et al., 1992). However, self-regulated academic behaviors, such as active participation, are contingent on motivational as well as cognitive factors (Zimmerman, 1986). Even young children with limited knowledge may participate actively if interested or motivated to do so. Specifically, children's sense of competence and control has been related to their level of engagement in classroom activities (Mischel & Shoda, 1995; Schunk, 1996). In addition, children have been found to be more involved in academic tasks when there is a perceived need or reason to exert effort, such as recognition of the value of learning or the desire to reap educational benefits (Sansone, Wiebe, & Morgan, 1999). Early interests that are better matched to the activities in school may relate to increased participation in classroom discussions, making interactions with information familiar and appreciable and more likely to be judged by children as within their competencies. However, during select interest-driven activities, children access, interpret, and construct understanding of information within a variety of categories, and certain types of activities that differ in focus (material or social), referent (property or process), and transformative potential may direct attention to particular kinds of information. As a result, even if children do not differ in how frequently they participate in classroom discussions, depending on their early interest orientations, they still may vary in the specific types of information they contribute to (or elicit from) discussions, such as task- or topic-relevant information, process or strategy information, or performance or personally relevant information.

### *Pursuits of Information*

Although children must attend to teacher instruction in order to learn, children also benefit from self-initiated pursuits of information through asking questions or soliciting assistance (Newman & Goldin, 1990; Ryan & Pintrich, 1997; van der Meij, 1990). The information that learners generate for themselves may enhance or update knowledge, introduce or inform procedures and strategies, or provide feedback related to performance expectations and suc-

cess (Resnick, 1987; Schneider & Pressley, 1997; Tobias, 1994). Factors that are related to whether children pursue information include their prior forms of knowledge (van der Meij, 1990) or achievement (Newman & Goldin, 1990), perceptions of their cognitive and social competence (Newman, 1990; Ryan & Pintrich, 1997), goals for learning (Butler, 1998; Newman, 1990; Ryan & Pintrich, 1997), and beliefs in the benefits of asking for help (Newman, 1990; Ryan & Pintrich, 1997). Experience within certain activities with either a material or social focus and property or process referent may render some forms of information more customary or familiar and provide encounters that make salient the relevance or usefulness of certain types of information. It seems plausible then to hypothesize that children's early interest-based activities in the home may influence the types of information they pursue in school, such as factual, task-process, or normative information.

### The Present Study

The present study rested on several assumptions: (a) interests affect what children pay attention to and encode, as well as how they allocate effort during activities aligned with the interest; (b) young children cultivate interests through their activities in the preschool years; and (c) different interests are associated with specific opportunities to develop specialized knowledge or specific schemas for engaging with the world. Subsequently, young children's patterns of involvement and interactions with particular forms of information in school may differ as a function of their interest histories. This study investigates these assumptions longitudinally, examining children's early childhood interest-based activities in the home and later variations in their contributions to academic discussions and pursuits of information in kindergarten. Given that variations in cognitive aptitude, temperament, and gender could mask the relations between children's interests and academic behaviors, we statistically controlled these variables in order to examine the unique contribution of children's interests to their patterns of participation and information pursuits in the classroom.

### Method

#### *Participants*

One hundred nine children (58 boys and 51 girls) who would enter kindergarten the next school year were recruited from a larger sample participating in a longitudinal study of early interest development. Children were between the ages of 4 years, 0 months and 4 years, 6 months at the onset of the study ( $M = 4$  years, 2 months) and between the ages of 4 years, 11 months and 6 years, 0 months at the time of entrance to kindergarten ( $M = 5$  years, 6 months). Children were recruited from two Midwest cities and neighboring communities.

The majority of children were White (88% White, 3% African American, 4% Asian American, 1% Latino, and 5% biracial) and from middle- to upper-middle-class families. The average number of years of parents' education was 16.03 ( $SD = 1.76$ ) for mothers and 16.51 ( $SD = 2.40$ ) for fathers. On average, the families had three children. Annual family income ranged from less than \$15,000 to more than \$100,000, with a mean of approximately



\$55,000 ( $SD = \sim \$35,000$ ). Children in the sample on average had spent approximately 18 hr ( $SD = 14.71$ , range = 0–52) per week in day care 1 year before school entry.

### Procedure

Data on children's personal characteristics and home activities collected at intake and throughout the 12 months prior to school entry were used for this study. During an initial laboratory visit, a battery of tests was administered to assess the child's receptive language skills, cognitive efficiency, and short-term memory. During this initial laboratory visit, parents provided basic child and family demographic background information. Parents also reported their child's activities in the home. Two months after the baseline assessment and continuing for 1 year, bimonthly interviews with parents were conducted to regularly update information about the child's activities. During a second laboratory visit at the beginning of the second year of the study (generally conducted 1–5 months before kindergarten began), the parents completed questionnaires about their child's temperament as well as their child's activities in the home.

Children were observed in their kindergarten classrooms so that aspects of academic engagement could be evaluated. Four doctoral students in educational psychology (blind to home activity data and to specific hypotheses) were trained during a 1-month period to use an observational coding system to assess the targeted academic behaviors. After interrater reliability of 90% was reached on practice tapes, observations in the classroom began. To monitor coder consistency, we conducted frequent reliability checks using a second coder throughout the study. Each research assistant observed and coded behaviors for approximately one quarter of the sample. All data for a particular child were collected by the same research assistant.

Classroom observations began in the first month of school and were complete by the end of April. The observer was seated in the classroom, close enough to view the child's work and listen to verbalizations. Throughout 5-min observation intervals, the observer tallied each occurrence of target academic behaviors. An audio cassette player with headset and an audiotaped cue were used to signal completion of the 5-min observation interval. At the conclusion of the 5-min observation interval, the observer counted tallies and recorded total frequencies for each of the academic behavior categories. Eighteen 5-min intervals of data for each child were collected during academic work in kindergarten: six during independent work, six during small-group work, and six during teacher-directed instruction, for a total of 90 min (30 min during independent work, 30 min during small-group work, and 30 min during teacher-directed activities). The observer began 5-min intervals of data collection only when academic activities occurred in an appropriate learning context. No more than six 5-min intervals of data were collected for each child during a single classroom visit. Consequently, for each child, data were collected during at least three class visits at intervals approximately equally spaced throughout the school year.

### Measures

**Child's temperament.** Mothers or fathers completed scales from the Child Behavior Questionnaire (Goldsmith & Rothbart,

1991), a widely used measure of child temperament. Six scales were used: inhibitory–control, activity level, attention, approach–anticipation, shyness, and intensity of emotion. Parents rated 58 items (8 from the attention scale and 10 items from each of the other five scales) on a 7-point Likert scale (1 = *low*, 7 = *high*). To create scores for each of the scales, we reverse coded and averaged scale items when necessary. For each of the scales, a high score indicates a high level of the temperament characteristic assessed. We calculated composite control and emotional responsiveness scores used as covariates in the principal analyses of the study by reverse coding when necessary and averaging the appropriate scales. The inhibitory–control or restraint, attention, and activity level (reversed) scores were used to create the control composite, a measure of the child's general self-control abilities ( $\alpha = .74$ ). Approach–anticipation, shyness (reverse coded), and intensity of emotion scores were used to create the emotional responsiveness composite, a measure of the child's receptiveness to the environment and extroversion ( $\alpha = .71$ ).

**Cognitive ability.** The child's cognitive ability was assessed with a battery of tests to measure three principal areas: language, analytic skill, and working memory. Language skills (receptive vocabulary) were assessed using the Peabody Picture Vocabulary Test (Dunn & Dunn, 1997). Analytic skill was assessed using accuracy scores from the Kansas Reflection–Impulsivity Scale for Preschoolers (KRISP; Wright, 1971). The KRISP was designed to assess the extent to which a child selectively attends to perceptual details and is systematic in the visual processing of pictures. Verbal working memory was assessed using the Working Memory–Sentences Test (Siegel & Ryan, 1989). We calculated a composite cognitive skill score for use as a covariate in the principal analyses by standardizing and averaging the child's scores on each of the three cognitive measures ( $\alpha = .69$ ). Higher or lower composite scores provide a general estimate of the cognitive resources available to the child (high scores indicate high cognitive resources).

**Children's interest-based activities in the home.** A bimonthly parent report was used to gather information about the child's activities in the home. We coded parent responses to the following four questions: (a) "What does your child prefer to do during his/her free play time?" (b) "What kinds of toys or objects (bought in stores as well as found in nature) does your child like to play with the most?" (c) "Are there any special pretend themes that seem to reoccur in your child's play?" and (d) "If your child had 1 hr to play anything s/he wanted, what do you think s/he would prefer?" There were no restrictions on number of activities parents could mention.

Information from parents' open-ended responses to questions concerning their child's activities across the year was used to create a profile of scores for each child. First, each child activity reported by the parent was categorized into one of eight major activity types derived from previous research on young children's play (Bergen, 1988; Saracho, 1995; Smilansky, 1968): topic-centered exploration, construction or mechanical activity, games with rules, physical activity and/or sports, literacy, art and/or music, fantasy, and sociodramatic activity (see Appendix). We calculated the percentage of activities within each of the types by dividing the number of reported activities aligned with that type by the total number of reported activities for the year. Final activity profile scores represent the percentage (relative frequency) of



reported activities in each type during the year. To assess inter-coder agreement for each activity type, two investigators double-coded 25% of the reported activities independently. A version of Cohen's kappa appropriate for frequency data (Bartko & Carpenter, 1976; Berry & Mielke, 1988) was calculated to evaluate interrater agreement for each activity type score in the profile. Kappas ranged from .83 to 1.00.

There was considerable variability in the play experiences of individual children. On average, a young child participated in about 16 different activities as part of routine play in the home across the year ( $M = 15.65$ ,  $SD = 4.44$ , range = 9–27). The frequencies of activities in this sample were similar to those reported from other large studies of 4- to 6-year-olds' play interests (Macari, Simock & DeLoache, 2003; Simcock, Macari, & DeLoache, 2002). Although the children engaged in a wide array of activities in their homes, their play tended to be narrowly distributed across the major domains. About one third of the children were engaged in activities distributed across two play domains, about one third were engaged in activities across three play areas, and one third of the children were engaged in activities across four domains. In line with characteristics of play behavior (property vs. process referent, material vs. social focus, structure vs. transformative potential) described by Renninger (1984, 1989) in previous research, patterns in the children's profiles were suggestive of underlying interests in particular forms and features of activity, seemingly indicators of conceptual, procedural, creative, or socially oriented interests. Children were assigned to one of four interest groups on the basis of evaluation of the defining characteristics of the activity types and combination of types prominent in their personal activity profiles (Table 1).<sup>1</sup>

Activity with a property referent and transformative potential distinguished the play of some children. These children, assigned to the *conceptual* interest group, spent the largest proportion of their activity in object or topic-centered explorations (63%). Their activities were characterized by interactions with materials with a focus on topical concepts (e.g., dinosaurs, astronomy) and the acquisition of domain-relevant information. Process-based systematic or programmatic activity with a material focus distinguished the play of other children. These children, assigned to the *procedural* interest group, spent large portions of their activity in construction or mechanical activities (23%), games with rules (16%), and sports (19%). These activities were characterized by interactions in various processes with materials that had inherent structure or prescribed uses, such as blocks, puzzles, Legos, and erector sets, or predetermined goals and routines, such as board games, card games, and sports. Activity with a process referent, material focus, and transformative potential distinguished the play of another group of children. These children, assigned to the *creative* interest group, spent large portions of their activity in music or art (40%), fantasy (17%), and literacy (11%) activities: Their activities involved interaction in various processes with flexible materials that could be used innovatively (i.e., paints, crayons, clay, glue, scissors, paper and pencil, other craft objects, and musical instruments) or contexts in which the child could enter imaginary worlds (i.e., books, other media, fantasy play). Activity with inherent structure and an interpersonal focus distinguished the play of a final group of children. Children assigned to the *socially oriented* interest group spent the largest proportion of their activity in sociodramatic activities (58%). Their activities involved interac-

Table 1  
*Average Proportion (Relative Frequency) of Activity in Each Type of Play for Children in the Four Interest Groups*

Play type	Interest group			
	Conceptual ( <i>n</i> = 27)	Procedural ( <i>n</i> = 30)	Creative ( <i>n</i> = 24)	Social ( <i>n</i> = 28)
Topical exploration				
<i>M</i>	.63 <sup>a</sup>	.12	.03	.02
<i>SD</i>	.32	.27	.06	.04
Construction/mechanical				
<i>M</i>	.05	.23 <sup>a</sup>	.04	.06
<i>SD</i>	.10	.27	.15	.02
Games with rules				
<i>M</i>	.06	.16 <sup>a</sup>	.02	.07
<i>SD</i>	.11	.18	.05	.14
Physical/outdoor/sports				
<i>M</i>	.03	.19 <sup>a</sup>	.05	.10
<i>SD</i>	.07	.28	.11	.21
Literacy				
<i>M</i>	.02	.06	.11 <sup>a</sup>	.04
<i>SD</i>	.04	.14	.18	.10
Fantasy				
<i>M</i>	.05	.07	.17 <sup>a</sup>	.08
<i>SD</i>	.05	.10	.04	.03
Art or music				
<i>M</i>	.03	.03	.40 <sup>a</sup>	.08
<i>SD</i>	.07	.07	.37	.17
Sociodramatic				
<i>M</i>	.15	.13	.22	.58 <sup>a</sup>
<i>SD</i>	.24	.20	.31	.34

<sup>a</sup>Types of play distinguishing the activity of each interest group from all other interest groups on the basis of the results of Scheffé complex comparisons.

tions with toys that served as props for various types of social reconstructive role play (i.e., relationships, social roles, occupational themes).

We conducted two sets of analyses in order to verify that relatively homogeneous groups could be differentiated from patterns in the children's profiles of activities across the eight play types. First, we performed complex contrasts using the Scheffé multiple comparison procedure to determine if inimitable domains of activity differentiated each group from all other interest groups (Table 1). The percentage of topic-centered activities reported was significantly higher for children in the conceptual group than for all the other children,  $t(105) = 11.75$ ,  $p < .001$ . The percentages of reported involvement with construction or mechanical activities, games with rules play, and sports were significantly higher for children in the procedural interest group than for the other children,  $t(105) = 4.71$ ,  $p < .001$ ;  $t(105) = 2.25$ ,  $p = .02$ ; and  $t(105) = 2.17$ ,  $p = .03$ , respectively. The percentages of music and

<sup>1</sup> The purpose of the original larger study from which participants for the current study were drawn was to investigate expertise in early childhood. As a consequence, the current study may include more children with conceptual interests than one might expect to see in the general population of preschool-aged children. One might expect the other three interest types to be fairly evenly distributed; however, previous research indicates that only about one third of preschool children develop conceptual interests (Johnson et al., 2004).



art activities, fantasy play, and literacy activities reported was significantly higher for children in the creative interest group than for all the other children,  $t(105) = 7.69, p < .001$ ;  $t(105) = 5.49, p < .001$ ; and  $t(105) = 2.59, p = .01$ , respectively. The percentage of dramatic activities reported was significantly higher for children in the socially oriented interest group than for all the other children,  $t(105) = 6.32, p < .001$ .

Next, a k-means cluster analysis of the eight activity types (requesting four groups) was performed. Group assignments generated in the cluster analysis provided support for group assignments made on the basis of qualitative analysis (92% agreement). For the cases in dispute ( $n = 9$ ), final group placements were made on the basis of qualitative analysis rather than the mathematical solution (Anderberg, 1973).<sup>2</sup> Given support for our hypothesis that children's patterns of activity reflect underlying comprehensive interests in particular types of activities, we were able to proceed in the investigation of the study's principal hypothesis: Children's early interests expressed through play may be related to differences in children's academic behaviors in school.

*Children's academic behaviors in kindergarten.* We assessed the children's academic contributions and pursuits in the kindergarten classroom using an observational coding system adapted from an instrument developed in previous studies of children's academic self-regulation in third-grade (Stright, Neitzel, Sears, & Hoke-Sinex, 2001) and in kindergarten (Neitzel & Stright, 2003). The original system captured information only about children's frequency of contributions to discussions and information pursuits. Refinements were made to facilitate more detailed assessments through specification of the types of information contributed and pursued in the classroom. Because the likelihood of classroom environmental effects cannot be ignored, observations of each child were divided equally in three academic contexts—teacher-directed, small-group work, and independent work—in an effort to control for these effects and ensure a representative sampling of all the children's behaviors.

Frequency counts of three types of information contributions to class discussions were made: providing elaborations or extensions, sharing suggestions or rationales, and generating connections or associations. Instances of the child sharing topic- or project-related facts or details were counted as elaborations—extensions. Instances of the child offering possible solutions, explanations or rationales, or suggestions for strategies to try were counted as suggestions—rationales. Instances of the child making ties between the topic of discussion and personal experiences or relating that topic to other issues were counted as connections—associations. We calculated final scores for each information contribution type by adding the frequency counts for each behavior from all observation periods.

We assessed the child's pursuits of information in the classroom using frequency counts of instances in which the child sought information in the classroom from the teacher, a peer, or other resources in the classroom. The child's pursuits of three information types were coded as additional information, task-process information, or normative information. Instances of the child seeking objective, factual information or requesting more details about a topic of study were counted as pursuits of additional information. Instances of the child seeking task or process (mastery) information, which includes strategies, rationales, and explanations of procedures or how and why things work, were counted as task-process information pursuits. Instances of the child seeking infor-

mation for the purpose of making social comparisons, evaluating performance in relation to peers, or determining relative standing ("Did I do it the way it's suppose to be?" "How did everyone else do it?" "Is mine as good as his?") were counted as normative information pursuits. We calculated final scores for each type of information pursuit by adding the frequency counts for each behavior from all observation periods.

Previous research has shown high correlations between parent instructional support behaviors in the home and observations of children's subsequent classroom involvement behaviors (Neitzel & Stright, 2003; Stright et al., 2001) and between observations and teacher ratings of these classroom behaviors (Stright & Neitzel, 2003). To assess intercoder agreement for the academic information contributions and pursuits in this study, Carin Neitzel observed jointly with each of four trained graduate research assistants and independently coded 25% of the classroom observations done by each of the research assistants. The observations with each research assistant were conducted at intervals equally spaced across the data collection period. For each academic information contribution and pursuit type, we calculated intercoder agreement using a variation of Cohen's kappa appropriate for frequency data (Scott & Hattfield, 1985). Cohen's kappas ranged from .82 to .95 for types of contributions to class discussions and .83 to .97 for types of information pursued.

## Results

First, preliminary analyses were conducted in order to provide descriptive information about the children and their interest-based activities as well as their academic behaviors and to check the assumptions associated with analyses of variance with covariates. In addition to the initial analyses reported here, the distributions of each variable in each of the four interest groups also were examined statistically and graphically. All assumptions of normality, homogeneity of variance, and linearity were upheld. Next, we conducted two split-plot factorial analyses of covariance (ANCOVAs) to examine differences among the interest groups in their academic behaviors.

### *Children's Personal Characteristics and Interests*

Frequencies, ranges, means, and standard deviations were calculated for each of the child characteristics variables; and analyses were conducted to assess gender, cognitive aptitude, and temperament differences across the four interest groups. There were statistically significant differences in interest group membership based on gender,  $\chi^2(3, N = 109) = 29.49, p < .001$ , Cramer's  $V = .52$  (see Table 2). Boys more often than girls had concept-oriented interests (proportions of .36 and .12, respectively) and procedural interests (proportions of .38 and .16, respectively); girls more often than boys had creative interests (proportions of .37 and .09, respectively) and socially oriented interests (proportions of .35 and .17, respectively).

<sup>2</sup> To ensure that this grouping decision did not alter the findings, we conducted all group comparison analyses using both grouping methods. There were not significant differences in the outcomes of these analyses, and no changes in the interpretations or conclusions drawn as a result of the grouping decision.



Table 2  
*Play Interest Group Membership by Gender*

Gender	Interest group			
	Conceptual ( <i>n</i> = 27)	Procedural ( <i>n</i> = 30)	Creative ( <i>n</i> = 24)	Social ( <i>n</i> = 28)
Boys	21	22	5	10
Girls	6	8	19	18

There also were statistically significant differences among the four interest groups in children's personal characteristics, Wilk's  $\lambda = .70$ ,  $F(9, 250) = 7.52$ ,  $p < .001$ , partial  $\eta^2 = .18$  (Table 3). Specifically, there were differences among the interest groups in cognitive skill,  $F(3, 105) = 7.21$ ,  $p < .001$ , partial  $\eta^2 = .17$ ; control,  $F(3, 105) = 5.04$ ,  $p = .003$ , partial  $\eta^2 = .13$ ; and emotional responsiveness,  $F(3, 105) = 3.82$ ,  $p = .01$ , partial  $\eta^2 = .10$ . Follow-up tests revealed that on average, the children in the concept-oriented interest group had higher cognitive skills than the children in the creative,  $t(49) = 3.11$ ,  $SE = 0.18$ ,  $p = .002$ , and procedural,  $t(55) = 2.35$ ,  $SE = 0.17$ ,  $p = .02$ , interest groups. Typically, children in the creative and procedural interest groups had higher levels of control than children in the socially oriented interest group,  $t(50) = 3.15$ ,  $SE = 0.17$ ,  $p = .002$ , and  $t(56) = 3.51$ ,  $SE = 0.16$ ,  $p = .001$ , respectively. On average, children in the concept-oriented and socially oriented interest groups were more emotionally responsive than children in the procedural interest group,  $t(55) = 2.76$ ,  $SE = 0.19$ ,  $p = .007$ , and  $t(56) = 2.36$ ,  $SE = .18$ ,  $p = .02$ , respectively. There were no systematic differences between boys and girls in cognitive skill or temperament characteristics.

#### *Children's Academic Behaviors in the Kindergarten Classroom*

Ranges, means, and standard deviations were calculated for each of the academic behaviors observed during the 90 min of

classroom observation (Table 4). In addition, we conducted bivariate correlations to examine the relations among children's different academic behaviors. Children who frequently pursued additional information in the classroom were likely to add elaborations—extensions during class discussions ( $r = .65$ ,  $p < .001$ ) and were less likely to make associations—connections during class discussions ( $r = -.30$ ,  $p = .002$ ) or to pursue normative information ( $r = -.24$ ,  $p = .01$ ). Children who frequently pursued task-process information in the classroom were likely to offer suggestions—rationales during class discussions ( $r = .49$ ,  $p < .001$ ) and to pursue normative information ( $r = .19$ ,  $p = .05$ ). Children who frequently made connections—associations were less likely to elaborate or extend class discussions ( $r = -.25$ ,  $p = .01$ ) but were more likely to offer suggestions—rationales during discussions ( $r = .24$ ,  $p = .01$ ) and to pursue normative information ( $r = .19$ ,  $p = .05$ ).

A series of multivariate analyses of variance (MANOVAs) were conducted to determine whether there were systematic differences between boys and girls in their frequency of academic behaviors in the classroom. In general, boys more often than girls contributed to class discussions, Wilk's  $\lambda = .83$ ,  $F(3, 105) = 7.17$ ,  $p < .001$ , partial  $\eta^2 = .17$ , and pursued information in class, Wilk's  $\lambda = .77$ ,  $F(3, 105) = 10.55$ ,  $p < .001$ , partial  $\eta^2 = .23$ . Specifically, follow-up analyses of variance (ANOVAs) on each behavior revealed that boys more often than girls provided elaborations—extensions,  $F(1, 107) = 15.77$ ,  $p < .001$ , partial  $\eta^2 = .13$ , and suggestions—rationales,  $F(1, 107) = 8.71$ ,  $p = .004$ , partial  $\eta^2 = .08$ . Girls did not differ from boys in their frequency of connections—associations made during class discussions. Boys more often than girls pursued additional information,  $F(1, 107) = 13.68$ ,  $p < .001$ , partial  $\eta^2 = .11$ , and task-process information,  $F(1, 107) = 8.96$ ,  $p = .003$ , partial  $\eta^2 = .08$ . Girls did not differ from boys in their normative information pursuits.

The relations among child characteristics and academic behaviors were examined through bivariate correlations. Children who had higher cognitive skills were likely to share more elaborations—extensions in class discussions ( $r = .26$ ,  $p = .01$ ) and pursue

Table 3  
*Means, Standard Deviations, and Ranges in Personal Characteristics of Children in Each Play Interest Group*

Personal characteristic	Interest group			
	Conceptual ( <i>n</i> = 27)	Procedural ( <i>n</i> = 30)	Creative ( <i>n</i> = 24)	Social ( <i>n</i> = 28)
Child cognitive skill				
<i>M</i>	0.34 <sub>a</sub>	−0.16 <sub>b,c</sub>	−0.31 <sub>c</sub>	0.26 <sub>a,b</sub>
<i>SD</i>	0.60	0.67	0.55	0.58
Range	−0.53–1.74	−1.12–1.58	−1.80–0.82	−0.97–1.03
Child control				
<i>M</i>	4.97 <sub>a,b</sub>	5.22 <sub>a</sub>	5.19 <sub>a</sub>	4.67 <sub>b</sub>
<i>SD</i>	0.39	0.60	0.54	0.77
Range	4.15–5.85	4.03–6.66	4.16–5.96	3.08–6.26
Child responsiveness				
<i>M</i>	4.86 <sub>a</sub>	4.36 <sub>b</sub>	4.59 <sub>a,b</sub>	4.80 <sub>a</sub>
<i>SD</i>	0.69	0.62	0.79	0.66
Range	3.82–5.08	3.36–5.48	3.76–6.48	3.11–5.41

*Note.* Means sharing a common subscript are not significantly different by the Dunn–Sidak multiple comparison test ( $p < .05$ ).

Table 4  
*Descriptive Statistics for Children's Academic Behaviors During 90 Min of Observation (N = 109)*

Child academic behaviors	<i>M</i>	<i>SD</i>	Range
Contribution to class discussions			
Elaboration-extension	3.82	3.20	0-13
Suggestion-rationale	5.06	3.73	0-14
Connection-association	4.47	3.27	0-12
Pursuit of information			
Additional information	2.10	1.98	0-9
Task-process information	3.63	2.77	0-13
Normative information	4.55	2.93	0-13

additional information at higher frequencies ( $r = .25, p = .01$ ). Children described as higher in control were more likely to offer suggestions-rationales during discussions ( $r = .23, p = .02$ ) and pursue task-process information in class ( $r = .19, p = .04$ ). Children described as higher in responsiveness were less likely to pursue task-process information ( $r = -.23, p = .02$ ); however, they were likely to pursue normative information more frequently ( $r = .19, p = .05$ ).

#### *Differences in Children's Academic Behaviors Related to Children's Interests*

We statistically controlled children's gender, cognitive skills, and temperament, characteristics that could compete to explain differences in children's academic behaviors by entering them as covariates in all analyses. Statistical removal of the variance in children's behaviors associated with their gender, cognitive skills, and temperament permitted an examination of the distinctive contribution of children's interests to understanding differences in their contributions to class discussions and information pursuits in kindergarten.<sup>3</sup>

**Contributions to discussions.** We conducted a split-plot factorial 4.3 analysis of covariance (ANCOVA) to examine children's participation in discussions. Children's interest type was the between-subjects factor (four levels), and children's type of contributions to class discussions was the within-subjects factor (three levels), with the effects of children's gender, cognitive skill, and temperament being controlled. However, gender was the only covariate with a significant unique effect on children's overall frequency of contributions to classroom discussions,  $F(1, 93) = 6.81, p = .01$ , partial  $\eta^2 = .07$ .

In the between-subjects analysis of variance, there were no differences among the interest groups in their overall frequency of participation in discussions in the classroom,  $F(3, 101) = 0.44, p = .73$ , partial  $\eta^2 = .01$ . In the within-subjects analysis, significant differences in frequency were found in the types of information that the children contributed to discussions in class,  $F(2, 100) = 4.96, p = .01$ , partial  $\eta^2 = .10$ . In general, the children contributed suggestions-rationales more often than elaborations-extensions. However, a significant interaction signaled differences in the types of information children contributed to discussions depending on their interests,  $F(6, 202) = 3.65, p = .002$ , partial  $\eta^2 = .10$  (see Figure 1).

To identify the differences in children's profiles of contributions to discussions, we conducted factorial profile (repeated measures)

analyses and post hoc comparisons for each group to examine differences in the types of contributions made within each interest group, controlling for the familywise error rate using Holm's sequential Bonferroni approach (Table 5; comparisons across each row). There were significant differences in the frequency of behaviors within the profiles of each interest group: conceptual,  $F(2, 52) = 5.76, p = .005$ , partial  $\eta^2 = .18$ ; procedural,  $F(2, 58) = 6.10, p = .004$ , partial  $\eta^2 = .17$ ; creative,  $F(2, 46) = 4.66, p = .01$ , partial  $\eta^2 = .17$ ; and socially oriented,  $F(2, 54) = 4.67, p = .01$ , partial  $\eta^2 = .15$ . Children with conceptual interests shared elaborations-extensions more often than they made connections-associations,  $t(26) = 2.93, SE = 0.96, p = .01$ . Children who had procedural interests shared suggestions-rationales more often than elaborations-extensions,  $t(29) = 3.36, SE = 0.91, p = .002$ , or connections-associations,  $t(29) = 2.36, SE = 0.87, p = .03$ . Children with creative interests made connections-associations and shared suggestions-rationales more often than elaborations-extensions,  $t(23) = 2.49, SE = 0.95, p = .02$ , and  $t(23) = 3.02, SE = 0.73, p = .01$ , respectively. Children with socially oriented interests made connections-associations more often than elaborations-extensions,  $t(27) = 2.55, SE = 0.83, p = .02$ .

In addition, a series of one-way ANCOVAs and Holm's sequential Bonferroni follow-up comparisons were conducted to assess differences among interest groups in the frequency of their three types of contributions to class discussions with the effects of children's gender, cognitive skill, and temperament controlled (Table 5; comparisons within each column). There were significant differences among groups in their elaborations-extensions,  $F(3, 101) = 3.98, p = .01$ , partial  $\eta^2 = .11$ ; suggestions-rationales,  $F(3, 101) = 2.95, p = .04$ , partial  $\eta^2 = .08$ ; and connections-associations,  $F(3, 101) = 2.61, p = .05$ ; partial  $\eta^2 = .07$ . Children with conceptual interests provided elaborations-extensions more than children whose interests were procedural,  $t(55) = 2.86, SE = 0.83, p = .01$ ; creative,  $t(49) = 2.37, SE = 0.98, p = .03$ ; or socially oriented,  $t(53) = 2.99, SE = 0.94, p = .01$ . Children who had procedural interests talked about suggestions-rationales in class discussions more often than children who had socially oriented interests,  $t(56) = 2.84, SE = 0.68, p = .01$ . Children who had creative or socially oriented interests made connections-associations more often than children who had conceptual interests,  $t(49) = 2.68, SE = 1.03, p = .03$ , and  $t(53) = 2.52, SE = 1.01, p = .03$ , respectively.

**Information pursuits.** To examine children's information pursuit behaviors, we conducted a split-plot factorial 4.3 ANCOVA. Again, children's interest type was the between-subjects factor (four levels) and the children's type of informational pursuits was the within-subjects factor (three levels), with the effects of children's gender, cognitive skill, and temperament controlled. However, gender was the only covariate with a significant unique relationship to children's overall frequency of information pursuits in the classroom,  $F(1, 93) = 3.88, p = .05$ , partial  $\eta^2 = .04$ .

In the between-subjects analysis of variance, there were no differences among the interest groups in their overall frequency of information pursuits in the kindergarten classroom,  $F(3, 101) =$

<sup>3</sup> It should be noted that the outcomes, interpretations, and conclusions drawn from initial analyses conducted without covariates included did not change with entry of the covariates.



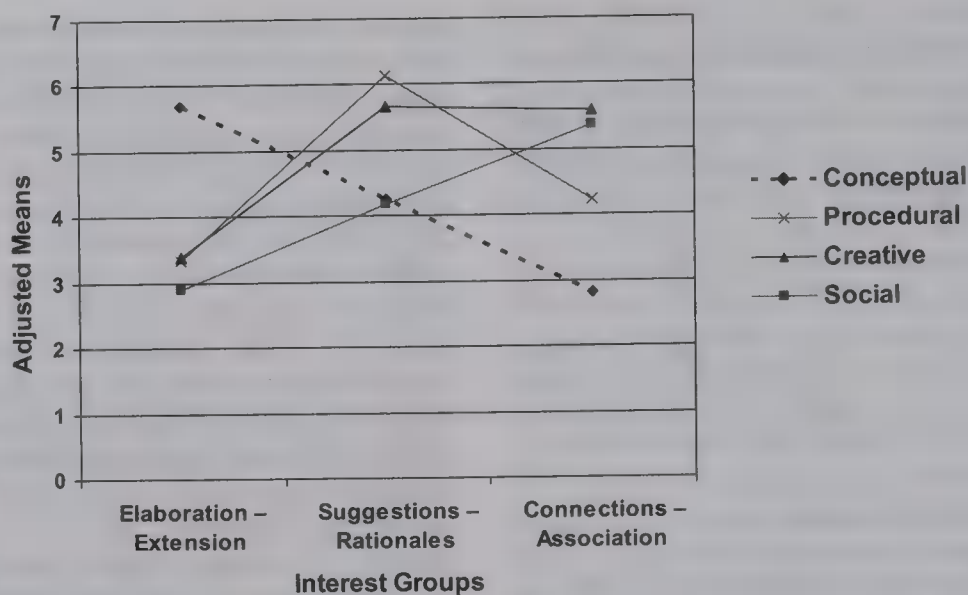


Figure 1. Plot of interactions between children's interests and types of contributions made to academic discussion in the kindergarten classroom.

1.12,  $p = .34$ , partial  $\eta^2 = .03$ . In the within-subjects analysis, significant differences in frequency were found in the types of information that the children pursued in the classroom,  $F(2, 100) = 26.25$ ,  $p < .001$ , partial  $\eta^2 = .34$ . In general, the children pursued normative information more often than additional and task-process information and pursued task-process information more often than additional information. However, a significant interaction signaled differences in the types of information the

children pursued depending on their interests,  $F(6, 202) = 8.81$ ,  $p < .001$ , partial  $\eta^2 = .21$  (see Figure 2).

To identify the differences in the children's profiles of information pursuits and the frequency of the types of information pursued within each of the interest groups, we conducted factorial profile (repeated measures) analyses and post hoc comparisons, controlling for familywise error rate using Holm's sequential Bonferroni approach (Table 6; comparisons across each row). Children who had conceptual interests did not differ in the frequency of their pursuits of the various types of information,  $F(2, 52) = 0.95$ ,  $p = .39$ , partial  $\eta^2 = .04$ , but there were significant differences in the frequency of behaviors within the profiles of children whose interests were procedural,  $F(2, 58) = 19.38$ ,  $p < .001$ , partial  $\eta^2 = .40$ ; creative,  $F(2, 46) = 18.42$ ,  $p < .001$ , partial  $\eta^2 = .45$ ; and socially oriented,  $F(2, 54) = 35.75$ ,  $p < .001$ , partial  $\eta^2 = .57$ . Children who had procedural interests sought task-process information and normative information more often than additional information,  $t(29) = 5.71$ ,  $SE = 0.65$ ,  $p < .001$ , and  $t(29) = 5.10$ ,  $SE = 0.74$ ,  $p < .001$ , respectively. Children with creative interests

Table 5

Post Hoc Comparisons Examining Differences in the Frequency of Types of Contributions to Class Discussion Within Each Interest Group (Rows) and Between Interest Groups For Each Type of Contribution to Class Discussions (Columns)

Interest group	Frequency of contributions to class discussion		
	Elaborations or extensions	Suggestions or rationales	Connections or associations
Conceptual ( $n = 27$ )			
<i>M</i>	5.70 <sub>a,c</sub>	4.25 <sub>a,b,c,d</sub>	2.82 <sub>b,d</sub>
<i>SE</i>	0.64	0.75	0.73
Procedural ( $n = 30$ )			
<i>M</i>	3.33 <sub>b,d</sub>	6.13 <sub>a,c</sub>	4.22 <sub>b,c,d</sub>
<i>SE</i>	0.64	0.79	0.73
Creative ( $n = 24$ )			
<i>M</i>	3.38 <sub>b,d</sub>	5.67 <sub>a,c,d</sub>	5.59 <sub>a,c</sub>
<i>SE</i>	0.75	0.92	0.86
Socially oriented ( $n = 28$ )			
<i>M</i>	2.89 <sub>b,d</sub>	4.19 <sub>a,b,d</sub>	5.36 <sub>a,c</sub>
<i>SE</i>	0.68	0.83	0.77

Note. Frequency of contributions = estimated means after model covariates (child characteristics: gender, cognitive skill, control, and emotional responsiveness) were controlled. For each interest group, means sharing a common subscript are not significantly different by the Holm's sequential Bonferroni comparison procedure ( $p < .05$ ). Subscripts a and b denote comparisons across each row (within-group comparisons). Subscripts c and d denote comparisons within each column (between-group comparisons).

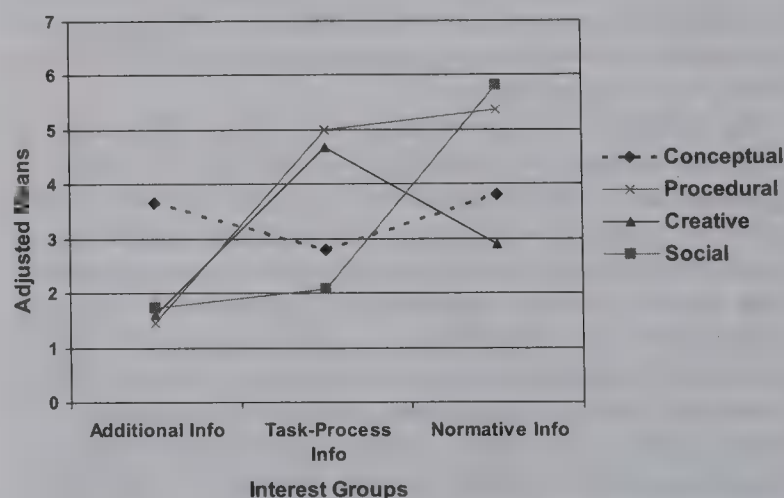


Figure 2. Plot of interactions between children's interests and the types of information pursued in the kindergarten classroom.

Table 6  
*Post Hoc Comparisons Examining Differences in the Frequency of Types of Information Pursuits Within Each Interest Group (Rows) and Between Interest Groups (Columns) for Each Type of Information Pursuit*

Interest group	Frequency of information pursuit		
	Additional information	Task-process information	Normative information
Conceptual ( <i>n</i> = 27)			
<i>M</i>	3.66 <sub>a,d</sub>	2.80 <sub>a,e,f</sub>	3.80 <sub>a,e</sub>
<i>SE</i>	0.36	0.56	0.61
Procedural ( <i>n</i> = 30)			
<i>M</i>	1.45 <sub>b,e</sub>	4.99 <sub>a,d</sub>	5.36 <sub>a,d</sub>
<i>SE</i>	0.36	0.56	0.61
Creative ( <i>n</i> = 24)			
<i>M</i>	1.62 <sub>c,e</sub>	4.67 <sub>a,d,e</sub>	2.89 <sub>b,e</sub>
<i>SE</i>	0.42	0.65	0.71
Socially oriented ( <i>n</i> = 28)			
<i>M</i>	1.72 <sub>b,e</sub>	2.08 <sub>b,f</sub>	5.82 <sub>a,d</sub>
<i>SE</i>	0.38	0.59	0.64

*Note.* Frequency of information pursuit = estimated means after model covariates (child characteristics: gender, cognitive skill, control, and emotional responsiveness) were controlled. For each interest group, means sharing a common subscript are not significantly different by the Holm's sequential Bonferroni comparison procedure ( $p < .05$ ). Subscripts a, b, and c denote comparisons across each row (within-group comparisons). Subscripts d, e, and f denote comparisons within each column (between-group comparisons).

sought more task-process information than normative,  $t(23) = 3.00$ ,  $SE = 0.40$ ,  $p = .01$ , or additional,  $t(23) = 6.32$ ,  $SE = 0.40$ ,  $p < .001$ , information and more normative than additional information,  $t(23) = 2.97$ ,  $SE = 0.44$ ,  $p = .01$ . Children who had socially oriented interests sought normative information more than task-process,  $t(27) = 6.80$ ,  $SE = 0.56$ ,  $p < .001$ , or additional,  $t(27) = 6.39$ ,  $SE = 0.69$ ,  $p < .001$ , information.

In addition, we conducted a series of ANCOVAs and follow-up comparisons using the Holm's sequential Bonferroni approach to assess differences among the interest groups in frequency of their pursuits of each information type with the effects of children's gender, cognitive skill, and temperament controlled (Table 6; comparisons within each column). There were significant differences among groups in their pursuits of additional information,  $F(3, 101) = 8.82$ ,  $p < .001$ , partial  $\eta^2 = .21$ ; task-process information,  $F(3, 101) = 5.92$ ,  $p = .001$ , partial  $\eta^2 = .15$ ; and normative information,  $F(3, 101) = 5.16$ ,  $p = .002$ , partial  $\eta^2 = .13$ . Children who had conceptual interests pursued additional information more often than children whose interests were procedural,  $t(55) = 4.70$ ,  $SE = 0.47$ ,  $p < .001$ ; creative,  $t(49) = 3.29$ ,  $SE = 0.62$ ,  $p = .001$ ; or socially oriented,  $t(53) = 3.46$ ,  $SE = 0.56$ ,  $p = .001$ . Children who had procedural interests sought task-process information more often than children with conceptual,  $t(55) = 3.04$ ,  $SE = 0.72$ ,  $p = .003$ , and socially oriented,  $t(56) = 3.23$ ,  $SE = 0.90$ ,  $p = .002$ , interests. Children who had creative interests also asked for task-process information more often than children with socially oriented interests,  $t(50) = 3.08$ ,  $SE = 0.84$ ,  $p = .003$ . Children with creative and conceptual interests sought normative information less often than children with procedural interests,  $t(52) = -2.68$ ,  $SE = 0.92$ ,  $p = .02$ , and  $t(55) = -2.05$ ,

$SE = 0.76$ ,  $p = .04$ , respectively, or socially oriented interests,  $t(50) = -3.26$ ,  $SE = 0.90$ ,  $p = .002$ , and  $t(49) = -2.14$ ,  $SE = 0.94$ ,  $p = .03$ , respectively.

## Discussion

This is the first known prospective investigation of the relations between the interest-related activities of a fairly large sample of preschoolers and their classroom behaviors at school entry. This study offers a detailed account of the complex relations between children's interest-based activities in the home during the early childhood years and subsequent forms of academic participation in kindergarten.

Children in four interest groups (conceptual, social, procedural, and creative) did not differ in their frequency of information contributions and pursuits in the kindergarten classroom. However, exploration of the trends both between and within interest groups revealed differences among the groups in the frequency of various types of information contributions and pursuits as well as unique profiles of academic participation behaviors within each group. There was remarkable consistency between the types of information that the children offered in discussions and pursued in the classroom and those that the children likely experienced or practiced through interest-based activity routines. Although we are cognizant of the fact that there are likely mediator variables we have not measured that could account for this consistency, it quite plausibly reflects a relationship between children's early engagement with particular types of information through interest-based activities in the home and later ways of interacting with information in school.

### *Children's Interests and Academic Information Contributions and Pursuits*

Children who had process-oriented interests (both procedural and creative) offered suggestions and provided rationales during class discussions and actively sought task-process information more frequently than did children with other interest orientations. The activities characteristic of process-oriented interests often involve trial and error, examination of interrelations among factors, and investigation of alternative approaches to problems. Therefore, these activities may support the acquisition of information about how tasks work and generation of general strategies, routines, or procedures for approaching tasks (Vandenberg, 1980, 1990). In these activities, the children may have learned to pay attention to specific aspects of tasks and problem-solving situations.

Children who had creative interests differed from children who had procedural interests in their information pursuits in ways consistent with the experiences they likely had in their early childhood activities. The children who had creative interests pursued normative information much less often than did children who had other interests and made connections and associations more often than the other children. Children with creative interests may be unconcerned with information for the purpose of social comparison because the activities that are characteristic of their interests usually permit exploration of possibilities without established standards for evaluation (Pulaski, 1970, 1973; Sutton-Smith, 1979). These children may be unaware of performance assessment



routines or the usefulness of information that can be gained, or they simply may derive their own benchmarks by which they measure their progress. Open-ended activities also are a forum for the generation of novel ideas (Clark, Griffing, & Johnson, 1989) and associations (Cole & Lavoie, 1985). The connections made by children in the creative interest group may be artifacts of a predilection for unique approaches or interpretations and, therefore, may represent cognitive elaborations (Klein & Kihlstrom, 1986; Symons & Johnson, 1997), or creative children may use connections as instruments to increase interest (Sansone et al., 1992; Wolters, 1999). We did not distinguish between connections and associations that were novel conceptualizations and those that were personal links between classroom topics of discussion and previous experiences or ways of understanding. In future research, measurements that allow teasing apart the particular function of young children's connections and associations could be beneficial.

Children who had socially oriented interests often made connections and associations during class discussions but offered suggestions and rationales as well. They most frequently pursued normative information. The activities of children with socially oriented interests involve the use of toys or props in the enactment of roles about life experiences focused on relationships or the roles or rules and structure of society (Jennings, 1975). These information-seeking behaviors may simply be a mirror of the children's socially oriented interests, or they may be a function of a lack of experience or practice with activities that support other forms of knowledge or interaction routines (Ericsson, Krampe, & Tesch-Romer, 1993). Future research that examines the metacognitive talk of children in class may assist researchers in isolating the most appropriate of these interpretations.

Children who had concept-oriented interests offered more facts or details to discussions, elaborating on and extending academic conversations, and made fewer connections and associations than most children. In contrast to children from other interest groups, they sought additional information about topics of study as often as other types of information. Children with concept-oriented interests typically engage in activities that involve interactions with materials, books, and other media, often with the goal of acquiring domain-specific information (Johnson, Alexander, Spencer, Leibham, & Neitzel, 2004). Consequently, these children may possess better-developed knowledge bases as well as more experience with the integration and organization of information (Chi & Koeske, 1983; Mervis, Pani, & Pani, 2003). In this way, the interest-relevant activities that are familiar and valued may be better matched to the information-dominated activities emphasized in classrooms (Super & Harkness, 1986) and may prepare the children for participation of this type in school.

### *Children's Academic Behaviors in Kindergarten*

In contrast to previous studies of academic participation behaviors that have focused on older children or adolescents (for a meta-analysis, see Winne & Perry, 2000), this study offers a rich analysis of the academic information exchanges and pursuits of young children in school. Although, the primary purpose of this study was to examine the differential effect of children's early interests on their subsequent forms of academic engagement, we would be remiss to ignore an opportunity to at least briefly examine this rare portrait of young children's general academic partici-

ipation patterns, a profile of behaviors touted in prior research as important to achievement in school (Normandeau & Guay, 1998).

In general, the young children in this study were actively involved in discussions and in the pursuit of information. An argument could be made that the academic behaviors observed might be specific to children from relatively well-educated families; however, similar results have been found in observational studies of kindergarteners from less-educated families (Neitzel & Stright, 2003). During academic discussions, these children shared suggestions and rationales most often, although they customarily made connections and associations as well. Elaborations or extensions were more rare. In information pursuits in the classroom, the children asked for normative information routinely, more than they asked for task-process information or additional information about topics.

Although involved, the children may not seem to have been particularly self-reliant. However, young learners who were not self-directed and motivated probably would not care about how they performed. Habitual pursuits of normative information may reflect young children's desire to understand the standards of operation and expectations for performance in the academic setting (Bandura, 1986) or to obtain information about themselves as learners (Ferne, 1989). Therefore, we do not consider these self-regulation deficiencies. Social comparisons allow young children to determine the appropriateness of their academic behaviors and appraise their performances (Butler & Winne, 1995; Ruble, 1987; Ruble & Flett, 1988). As a result, evaluations against normative standards also can promote self-efficacy (Schunk, 1987, 1996; Schunk & Hanson, 1989).

Similarly, connections and associations offered during class discussions may represent young children's attempts to locate their place within the academic arena or to establish personal relevance. Brophy (1999) has argued that for learning to occur, the material to be learned must be potentially meaningful. Personal references may be tactics used to establish links to new material (Masten & Coatsworth, 1998; Sansone & Smith, 2000; Spires & Donley, 1998). Children's connections and associations also may reflect strategic efforts to enhance learning through the integration of new information with existing knowledge (Symons & Johnson, 1997).

Often, solicitation of normative information and reliance on self-relevant ties have been described as precursors to or inhibitors of more independent or sophisticated forms of academic self-regulation (Schunk & Zimmerman, 1997; Wolters, 2004). These practices seem to conflict with the image of a mastery-oriented, motivated, and self-sufficient individual. However, these practices are probably valuable in situations that are unfamiliar or when understanding is fragile. Thus, these findings raise questions about whether different self-regulation strategies serve distinct functions at different times in development. This warrants attention in future research.

### *Implications for Parents and Teachers*

The information derived from this study has implications for both parents and teachers. Many parents worry about how best to support their young children as learners, and the kinds of experiences and structures they should put in place for optimal growth. At least in part, children's academic self-instruction behaviors seem to be cultivated and sustained in the course of routine



activity. Therefore, young children's early interests are not insignificant and may provide the fundamental context and processes for academic preparation. These findings suggest that opportunities to participate in a variety of experiences be made available to children in early childhood. Differences in children's early childhood experiences did not translate into differences in levels of engagement in school. All of the children in this study, regardless of their interest-based activities, participated equally in the classroom milieu.

In order to construct knowledge, learners must be actively engaged with information and activities available in classroom settings (McCaslin & Good, 1996). Participation is a critical tool for academic success (Schneider & Bjorklund, 1992). Reciprocally, information has a strong influence on cognitive engagement and forms of engagement with tasks (Winne, 1995). Teachers should be aware that although children with all types of experiences seem equally primed to participate in the classroom environment, differences in their histories of home activities mean that young children may come to school with different kinds of interaction preferences. Teachers need to be open to children's predictions for different types of information exchanges and queries and aware that each of the different types of participation may have important benefits for young learners.

These findings also serve as a reminder to teachers that young children rely on their feedback to learn about the values and behavioral expectations of school, and young children's initial presumptions about their competencies and capabilities are founded on information from external sources (Butler & Winne, 1995). Consequently, teachers must be careful about the expectations, values, goals, and attributions conveyed. Finally, teachers also should be cautious in the use of interest categories for understanding and evaluating the behaviors, needs, and potentials of young children. It certainly could be useful to capitalize on what is known about children's interest proclivities in the design of settings and instructional strategies to support young children as they transition to school, but it could be equally harmful if this information were used to narrowly define children. Although children may express preferences for particular kinds of experiences and interactions, children also engage in and enjoy a variety of activities. Recent research has shown that interests are susceptible to change, particularly as young children enter new social settings, encounter new opportunities, and expand their social networks (Wang, 2003).

### *Final Comments and Directions for Future Research*

Past research has linked childhood activities with cognitive development (Fisher, 1992), motivation (Christie & Johnsen, 1983), social competence (Creasey, 1998), and theory of mind (Lillard, 1998), which are all important tools for academic self-regulation. The fundamental assumption framing the present study was that early childhood activities serve as a context in which young children develop, express, and further cultivate interests. Another assumption guiding this study was that interest-based activities provide exposure to different experiences and opportunities to develop certain types of knowledge and interaction scripts or routines. As a result, children's early interests may lead to differences in their academic participation. The remarkable consistencies noted in children's interests and academic behaviors

sustain the plausibility of this explanation and compel further research.

We believe that individual interest could only be inferred from a child's persistence and consistency of engagement across various activity types; therefore, an effective investigation of our research hypotheses required a longitudinal tracking of children's home activities. Although converging observations of children's activities would have been ideal, this simply was not logistically feasible. Parents are apt to be better retrospective reporters of home activity histories than young children. It is true that parents could provide unreliable and biased accounts of their own children's activities. Research by Andre and Brown (1969) has suggested, however, that mothers perceive their children's interests relatively accurately. In addition, incomplete and erroneous reports only would have made it more difficult for systematic relationships between children's interests and academic engagement or differences among interests groups to be detected. The emergence of clear and theoretically consistent patterns of relations helps to reduce concerns associated with our reliance on parental report.

Additionally, in any attempt to examine children's academic behaviors, the likelihood of classroom setting effects cannot be ignored. We attempted to control for these effects by controlling for academic subject and instructional context. However, classrooms are complex contexts, and it would be difficult to identify—let alone to control for—all of the possible variations; therefore, it is quite likely that some error with the potential to mask the influence of early interests was introduced in the measurement of children's academic behaviors. However, meaningful patterns powerful enough to be noted over any "noise" in the data were uncovered. Still, future studies of these relations should include efforts to devise more sophisticated methods of controlling (or considering) important features of the classroom environment.

In this study, gender posed another methodological and conceptual challenge. Gender was related to children's interests as well as their academic behaviors, making it difficult to disentangle the influences of interest-based activity from those of gender. Consequently, despite our efforts to statistically control these effects, the question still may be raised about whether the differences observed in the present study in academic self-regulation behaviors are attributable to gender rather than to early interests. We maintain that patterns in the findings argue against gender as a primary explanation. If gender were the principal causal factor, the patterns in contributions to classroom discussions and pursuits would have been similar for the two interest groups predominately composed of boys (conceptual and procedural) and of girls (social and creative). This was not the case.

To the extent that children's personal characteristics such as gender, cognitive skill, and temperament were adequately controlled, this study provides a rare examination of the unique relations between children's interest-based activities in early childhood and academic self-instruction behaviors in school. Surprisingly few studies have controlled for individual difference factors that could confound such studies. However, more than a single score (or two) may be required to control for complex constructs such as cognitive skill and temperament. Sample size (although relatively large for a study of this type) imposed limits on the number of variables that could be included in the analyses. Replication with larger samples would enable the examination of more complex models. For example, a child's cognitive abilities, per-



sonal characteristics, and interests may work together to influence classroom behaviors. An alternative possibility is that a child's early childhood interests represent the juncture of personal and socialization factors and may be a more proximal source, mediator, or moderator of influence on behavior in new settings.

Interest in early childhood has been described as an artifact of social-contextual as well as interpersonal factors (Neitzel, Johnson, & Alexander, 2003). The material resources available, as well as procedures and structures instituted within the home, help shape children's expectancies, values, goals, and mental models for roles and action possibilities (Pintrich & Schrauben, 1992; Super & Harkness, 1986). Consequently, knowledge and interaction routines cultivated during early childhood activities may be fundamental to children's academic behaviors, or the same processes that support the emergence of particular interests in the home also may support the development of children's academic behaviors, or both may contribute uniquely. Additional research is needed to simultaneously examine the direct and indirect influences of important elements of the home environment and individual cognitive and personal resources of the child.

Further research also is needed to examine how long connections between early interests and academic behaviors endure. Findings in the present study suggest the influence of children's early interests appears relatively pervasive across the first year of school. However, as children spend increasingly more time in school, the many new and intriguing ideas and experiences encountered may exert greater influence. Thus, children are likely to move away from early predilections, making the influence of childhood interests less prominent over time. On the other hand, if early schemas for interactions are maintained, the influence asserted may persist. Investigation of these and other important questions will enhance theoretical models of academic self-regulation development in which the individual and social contexts each assume important roles.

## References

- Anderberg, M. (1973). *Cluster analysis for applications*. New York: Academic Press.
- Andre, V., & Brown, D. (1969). Children's interests and their mothers' perceptions of these interests. *Measurement and Evaluation in Guidance*, 2, 168-173.
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice-Hall.
- Bartko, J. J., & Carpenter, W. T. (1976). On the methods and theory of reliability. *Journal of Nervous and Mental Disease*, 163, 307-328.
- Baum, S. M., Renzulli, J. S., & Hebert, T. P. (1995). Reversing underachievement: Creative productivity as a systematic intervention. *Gifted Child Quarterly*, 39, 224-235.
- Bergen, D. (1988). *Play as a medium for learning and development: A handbook for theory and practice*. Portsmouth, NH: Heinemann Educational Books.
- Berry, K. J., & Mielke, P. W. (1988). A generalization of Cohen's kappa agreement measure to interval measurement and multiple raters. *Educational and Psychological Measurement*, 48, 921-933.
- Birch, S. H., & Ladd, G. W. (1998). Children's interpersonal behaviors and the teacher-child relationship. *Developmental Psychology*, 34, 934-946.
- Bjorklund, D. F., & Schneider, W. (1996). The interaction of knowledge, aptitude, and strategies in children's memory performance. *Advances in Child Development and Behavior*, 26, 59-89.
- Brophy, J. (1985). Interactions of male and female students with male and female teachers. In L. C. Wilkinson & C. B. Marrett (Eds.), *Gender influences in classroom interaction* (pp. 115-142). Orlando, FL: Academic Press.
- Brophy, J. (1999). Toward a model of the value aspects of motivation in education: Developing appreciation for particular learning domains and activities. *Educational Psychologist*, 34, 75-85.
- Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research*, 65, 245-281.
- Butler, R. (1998). Determinants of help seeking: Relations between perceived reasons for classroom help-avoidance and help-seeking behaviors in an experimental context. *Journal of Educational Psychology*, 90, 630-643.
- Chi, M. T., & Koeske, R. D. (1983). Network representation of a child's dinosaur knowledge. *Developmental Psychology*, 19, 29-39.
- Christie, J., & Johnsen, P. (1983). The role of play in social-intellectual development. *Review of Educational Research*, 53, 93-115.
- Clark, P. M., Griffing, P. S., & Johnson, L. G. (1989). Symbolic play and ideational fluency as aspects of the evolving divergent cognitive style in young children. *Early Child Development and Care*, 51, 77-88.
- Cole, D., & Lavoie, J. C. (1985). Fantasy play and related cognitive development in 2- to 6-year-olds. *Developmental Psychology*, 21, 233-240.
- Creasey, G. L. (1998). Play and social competence. In O. N. Saracho & B. Spodek (Eds.), *Multiple perspectives on play in early childhood education*. Albany, NY: State University of New York Press.
- Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Service.
- Ericsson, K. A., Krampe, R., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100, 363-406.
- Ericsson, K. A., & Smith, J. (1991). Prospects and limits of the empirical study of expertise: An introduction. In K. A. Ericsson & J. Smith (Eds.), *Toward a general theory of expertise: Prospects and limits* (pp. 1-13). New York: Cambridge University Press.
- Fernie, D. E. (1989). Becoming a student: Messages from first settings. *Theory Into Practice*, 27, 3-10.
- Fink, B. (1994). Interest and exploration: Exploratory action in the context of interest genesis. In H. Keller, K. Schneider, & B. Henderson (Eds.), *Curiosity and exploration* (pp. 100-120). New York: Springer-Verlag.
- Fisher, E. P. (1992). The impact of play on development: A meta-analysis. *Play and Culture*, 5, 159-181.
- Gati, I. (1991). The structure of vocational interests. *Psychological Bulletin*, 109, 309-324.
- Goldsmith, H. H., & Rothbart, M. K. (1991). Contemporary instruments for assessing early temperament by questionnaire and in the laboratory. In A. Angleitner & J. Strelau (Eds.), *Explorations in temperament* (pp. 249-272). New York: Plenum.
- Hidi, S. (1990). Interest and its contribution as a mental resource for learning. *Review of Educational Research*, 60, 549-571.
- Hidi, S., & Anderson, V. (1992). Situational interest and its impact on reading and expository writing. In K. A. Renninger, S. Hidi & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 215-238). Hillsdale, NJ: Erlbaum.
- Hidi, S., Berndorff, D., & Ainley, M. (2002). Children's argument writing, interest, and self-efficacy: An intervention study. *Learning and Instruction*, 12, 429-446.
- Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research*, 70, 151-180.
- Hidi, S., & McLaren, J. (1991). Motivational factors and writing: The role of topic interestingness. *European Journal of Psychology of Education*, 6, 187-197.



- Holmes, R. M. (1991). Categories of play: A kindergartner's view. *Play and Culture*, 4, 43–50.
- Jennings, K. (1975). People versus object orientation, social behavior, and intellectual abilities in preschool children. *Developmental Psychology*, 11, 511–519.
- Johnson, J. E., Christie, J. F., & Yawkey, T. D. (1987). *Play and early childhood development*. Glenview, IL: Scott, Foresman, & Co.
- Johnson, K. E., Alexander, J. M., Spencer, S., Leibham, M. E., & Neitzel, C. (2004). Factors associated with the early emergence of intense interests within conceptual domains. *Cognitive Development*, 19, 325–343.
- Jovanovic, J., & King, S. (1998). Boys and girls in the performance-based science classroom: Who's doing the performing? *American Educational Research Journal*, 35, 477–496.
- Klein, S. B., & Kihlstrom, J. F. (1986). Elaboration, organization, and the self-reference effect in memory. *Journal of Experimental Psychology: General*, 115, 26–38.
- Krapp, A. (2002). Structural and dynamic aspects of interest development: Theoretical considerations from an ontogenetic perspective. *Learning and Instruction*, 12, 383–409.
- Krapp, A., & Fink, B. (1992). The development and function of interests during the critical transition from home to preschool. In K. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 397–429). Hillsdale, NJ: Erlbaum.
- Krapp, A., Hidi, S., & Renninger, K. A. (1992). Interest, learning, and development. In K. A. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 3–25). Hillsdale, NJ: Erlbaum.
- Lillard, A. S. (1998). Playing with a theory of mind. In O. N. Saracho & B. Spodek (Eds.), *Multiple perspectives on play in early childhood education* (pp. 11–33). Albany, NY: State University of New York Press.
- Macari, S., Simcock, G. F., & DeLoache, J. S. (2003, April). *Tea sets, tractors, and trains: Intense interests in young children*. Paper presented at the biennial meeting of the Society for Research in Child Development, Tampa, FL.
- Masten, A. S., & Coatsworth, J. D. (1998). The development of competence in favorable and unfavorable environments: Lessons from research on successful children. *American Psychologist*, 53, 205–220.
- McCaslin, M. M., & Good, T. (1996). The informal curriculum. In D. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 622–670). New York: Macmillan.
- Mervis, C. B., Pani, J. R., & Pani, A. M. (2003). Category formation and evolution: Transaction of child interest, background conceptual knowledge, linguistic knowledge, and adult input in the acquisition of lexical categories at the basic and subordinate levels. In D. H. Rakison & L. M. Oakes (Eds.), *Early category and concept development*. Oxford, UK: Oxford University Press.
- Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, 102, 246–268.
- Neitzel, C., Johnson, K. E., & Alexander, J. M. (2003, October). *When predisposition meets opportunity: The emergence of children's early childhood interests expressed through play in the home*. Paper presented at the biennial meeting of the Cognitive Development Society, Park City, UT.
- Neitzel, C., & Stright, A. D. (2003). Relations between mothers' scaffolding and children's academic self-regulation: Establishing a culture of self-regulatory competence. *Journal of Family Psychology*, 17, 147–159.
- Newman, R. S. (1990). Children's help-seeking in the classroom: The role of motivational factors and attitudes. *Journal of Educational Psychology*, 82, 71–80.
- Newman, R. S., & Goldin, L. (1990). Children's reluctance to seek help with schoolwork. *Journal of Educational Psychology*, 82, 92–100.
- Normandeau, S., & Guay, F. (1998). Preschool behavior and first-grade school achievement: The mediational role of cognitive self-control. *Journal of Educational Psychology*, 90, 111–121.
- Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, 37, 91–105.
- Pintrich, P., & Schrauben. (1992). Students' motivational beliefs and their cognitive engagement in classroom academic tasks. In D. H. Schunk & J. L. Meece (Eds.), *Student perceptions in the classroom* (pp. 149–183). Hillsdale, NJ: Erlbaum.
- Pressick-Kilborn, K., & Walker, R. (2002). The social construction of interest in a learning community. In D. McInerney & S. Van Etten (Eds.), *Sociocultural influences on motivation and learning* (pp. 153–182). Greenwich, CT: Information Age.
- Pulaski, M. A. (1970). Play as a function of toy structure and fantasy predisposition. *Child Development*, 41, 531–537.
- Pulaski, M. A. (1973). Toys and imaginative play. In J. L. Singer (Ed.), *The child world of make-believe*. New York: Academic Press.
- Renninger, K. A. (1984). Object-child relations: Implications for both learning and teaching. *Children's Environment Quarterly*, 1, 3–6.
- Renninger, K. A. (1989). Individual patterns in children's play interests. In L. T. Winegar (Ed.), *Social interaction and the development of children's understanding* (pp. 147–172). Norwood, NJ: Ablex.
- Renninger, K. A. (1990). Children's play interests, representation, and activity. In R. Fivush & J. Hudson (Eds.), *Knowing and remembering in young children* (Vol. III, pp. 127–165). Cambridge, MA: Cambridge University Press.
- Renninger, K. A. (1992). Individual interest and development: Implications for theory and practice. In K. Renninger, S. Hidi, & A. Krapp (Eds.), *The role of interest in learning and development* (pp. 361–395). Hillsdale, NJ: Erlbaum.
- Renninger, K. A. (2000). Individual interest and its implications for understanding intrinsic motivation. In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic and extrinsic motivation: The search for optimal motivation and performance* (pp. 375–407). San Diego, CA: Academic Press.
- Renninger, K. A., & Hidi, S. (2002). Student interest and achievement: Developmental issues raised by a case study. In A. Wigfield & J. Eccles (Eds.), *The development of achievement motivation* (pp. 173–195). San Diego, CA: Academic Press.
- Renninger, K. A., & Leckrone, T. G. (1991). Temperament and task engagement: Individual patterns in children's play. In L. Oppenheimer & J. Valsiner (Eds.), *The origin of action: Interdisciplinary and international perspectives* (pp. 205–238). New York: Verlag.
- Renninger, K. A., & Wozniak, R. H. (1985). Effect of interest on attentional shift, recognition, and recall in young children. *Developmental Psychology*, 21, 624–634.
- Resnick, L. B. (1987). Constructing knowledge in school. In L. S. Liben (Ed.), *Development and learning: Conflict or congruence* (pp. 19–50). Hillsdale, NJ: Erlbaum.
- Ruble, D. N. (1987). The acquisition of self-knowledge: A socialization perspective. In N. Eisenberg (Ed.), *Contemporary topics in developmental psychology*. New York: Wiley.
- Ruble, D. N., & Flett, G. L. (1988). Conflicting goals in self-evaluative information seeking: Developmental and ability level analyses. *Child Development*, 59, 97–100.
- Ryan, A. M., & Pintrich, P. R. (1997). "Should I ask for help?" The role of motivation and attitudes in adolescents' help seeking in math class. *Journal of Educational Psychology*, 89, 329–341.
- Sansone, C., & Smith, J. L. (2000). The "how" of goal pursuit: Interest and self-regulation. *Psychological Inquiry*, 11, 306–309.



- Sansone, C., Weir, C., Harpster, L., & Morgan, C. (1992). Once a boring task, always a boring task? Interest as a self-regulatory mechanism. *Journal of Personality and Social Psychology*, 63, 379-390.
- Sansone, C., Wiebe, D. J., & Morgan, C. (1999). Self-regulating interest: The moderating role of hardiness and conscientiousness. *Journal of Personality*, 67, 701-733.
- Saracho, O. (1995). Relationship between young children's cognitive style and their play. *Early Childhood Development and Care*, 113, 77-84.
- Schiefele, U. (1990). The influence of topic interest, prior knowledge, and cognitive capabilities on text comprehension. In J. M. Pieters, K. Breuer, & P. R. J. Simons (Eds.), *Learning environments* (pp. 323-337). New York: Springer-Verlag.
- Schiefele, U. (2001). The role of interest in motivation and learning. In J. M. Collis & S. Messick (Eds.), *Intelligence and personality: Bridging the gap in theory and measurement*. (pp. 163-193). Mahwah, NJ: Erlbaum.
- Schneider, W., & Bjorklund, D. F. (1992). Expertise, aptitude, and strategic remembering. *Child Development*, 63, 461-473.
- Schneider, W., & Pressley, M. (1997). *Memory development between 2 and 20*. Hillsdale, NJ: Erlbaum.
- Schunk, D. H. (1987). Peer models and children's behavioral change. *Review of Educational Research*, 57, 149-174.
- Schunk, D. H. (1996). Goal and self-evaluative influences during children's cognitive skill learning. *American Educational Research Journal*, 33, 359-382.
- Schunk, D. H., & Hanson, A. R. (1989). Influence of peer-model attributes on children's beliefs and learning. *Journal of Educational Psychology*, 81, 431-434.
- Schunk, D. H., & Zimmerman, B. J. (1997). Social origins of self-regulatory competence. *Educational Psychologist*, 32, 195-208.
- Scott, M. M., & Hattfield, J. G. (1985). Problems of analyst and observer agreement in naturalistic narrative data. *Journal of Educational Measurement*, 22, 207-218.
- Sigel, I. E. (1993). The centrality of a distancing model for the development of representational competence. In R. R. Cocking & K. A. Renninger (Eds.), *The development of meaning of psychological distance* (pp. 141-158). Hillsdale, NJ: Erlbaum.
- Siegel, L. S., & Ryan, E. B. (1989). The development of working memory in normally achieving and subtypes of learning disabled children. *Child Development*, 60, 973-980.
- Simcock, G., Macari, S., & DeLoache, J. (2002, April). *Blenders, brushes, and balls: Intense interests in very young children*. Presented at the International Conference on Infant Studies, Toronto, ON, Canada.
- Smilansky, S. (1968). *The effects of sociodramatic play on disadvantaged preschool children*. New York: Wiley.
- Spires, H. A., & Donley, J. (1998). Prior knowledge activation: Inducing engagement with informational texts. *Journal of Educational Psychology*, 90, 249-260.
- Stright, A. D., & Neitzel, C. (2003). Beyond parenting: Coparenting and children's classroom adjustment. *International Journal of Behavioral Development*, 27, 31-40.
- Stright, A. D., Neitzel, C., Sears, K. G., & Hoke-Sinex, L. (2001). Instruction begins in the home: Relations between parental instruction and children's self-regulation in the classroom. *Journal of Educational Psychology*, 93, 456-466.
- Super, C., & Harkness, S. (1986). The developmental niche: A conceptualization at the interface of child and culture. *International Journal of Behavioral Development*, 9, 545-569.
- Sutton-Smith, B. (1979). Epilogue: Play as performance. In B. Sutton-Smith (Ed.), *Play and Learning* (pp. 295-320). New York: Gardner Press.
- Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin*, 121, 371-394.
- Tobias, S. (1994). Interest, prior knowledge, and learning. *Review of Educational Research*, 64, 37-54.
- Tracey, J. G. (1997). The structure of interests and self-efficacy expectations: An expanded examination of the spherical model of interests. *Journal of Counseling Psychology*, 44, 32-43.
- Vandenberg, B. (1980). Play, problem-solving, and creativity. In K. H. Rubin (Ed.), *Children's play* (pp. 49-68). San Francisco: Jossey-Bass.
- Vandenberg, B. (1990). Play and problem-solving: An elusive connection. *Merrill-Palmer Quarterly*, 36, 261-272.
- van der Meij, H. (1990). Question asking: To know that you do not know is not enough. *Journal of Educational Psychology*, 82, 505-512.
- Vermunt, J. D. (1996). Metacognitive, cognitive, and affective aspects of learning styles and strategies: A phenomenographic analysis. *Higher Education*, 31, 25-50.
- Vermunt, J. D. (1998). The regulation of constructive learning processes. *British Journal of Educational Psychology*, 68, 149-171.
- Wang, C. (2003, April). *The effects of curriculum support on preschool play interest focus and longevity*. Paper presented at the biennial meeting of the Society for Research in Child Development, Tampa, FL.
- Winne, P. H. (1995). Inherent details in self-regulated learning. *Educational Psychologist*, 30, 173-187.
- Winne, P. H., & Perry, N. E. (2000). Measuring self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeider (Eds.), *Handbook of self-regulation* (pp. 531-566). San Diego, CA: Academic Press.
- Woloshyn, V. E., Pressley, M., & Schneider, W. (1992). Elaborative-interrogation and prior-knowledge effects on learning of facts. *Journal of Educational Psychology*, 84, 115-124.
- Wolters, C. A. (1999). The relation between high school students' motivational regulation and their use of learning strategies, effort, and classroom performance. *Learning and Individual Differences*, 11, 281-299.
- Wolters, C. A. (2004). Advancing achievement goal theory: Using goal structures and goal orientations to predict students' motivation, cognition, and achievement. *Journal of Educational Psychology*, 96, 236-250.
- Wright, J. C. (1971). *Kansas Reflectivity Impulsivity Scale for Preschoolers (KRISP)*. St. Louis, MO: CEMREL.
- Zimmerman, B. J. (1986). Becoming a self-regulated learner: Which are the key subprocesses? *Contemporary Educational Psychology*, 11, 307-313.
- Zimmerman, B. J. (1989). A social cognitive view of self-regulated academic learning. *Journal of Educational Psychology*, 81, 329-339.
- Zimmerman, B. J., & Bandura, A. (1994). Impact of self-regulatory influences on writing course attainment. *American Educational Research Journal*, 31, 845-862.
- Zimmerman, B. J., & Martinez-Pons, M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. *American Educational Research Journal*, 23, 614-628.

## Appendix

## Interest Orientation Types; Toys or Materials; and Activities

Orientation/interest type	Toys or materials	Characteristic activities
Concept orientation—conceptual interest: Topic-centered exploration.	Interaction with toys or materials with a focus on object concepts and features. Characteristic topics: Dinosaurs, trains, insects, astronomy, geographic locations, ocean life, weather.	Seeking focal object/topical information (i.e., interactions with adults, books, media, or other information sources); questioning; organizing and managing information and object- or topic-relevant experiences.
Process orientation		
Procedural interest		
Construction or mechanical	Organizational toys designed for arrangement of parts: Puzzles, Cuisenaire rods, blocks, Legos, Tinker toys, peg boards, erector sets, models.	Examining interrelations of factors, one-to-one correspondence, and part-to-whole concepts; matching or grouping; producing; generating solutions; engaging in trial-and-error learning; investigating and evaluating alternate methods; practicing task process.
Games with rules	Board games, checkers, chess, tag, and hide and seek.	Practicing game-relevant skills, procedures/process information, strategies, and problem solving.
Physical or motor	Gross motor toys (e.g., swings, slides, bicycles, skates, and jump ropes) and sports equipment (e.g., bats, balls, and Frisbees).	Practicing relevant skills, procedures/process information, strategies, and problem solving.
Literacy (also may be creative, depending on activity focus)	Books, story books with tapes, reading and writing activities.	Focusing on learning the skill or process of reading.
Creative interest		
Art and music	Art supplies (e.g., paints, crayons, pencils and paper, markers, clay, play dough, scissors, glue, other craft materials); CDs or tapes; instruments; dance props; microphones.	Engaging in self-expression, creation, production, and generation of ideas or alternative methods; exploring material uses or transformations; processing information.
Fantasy	Materials used in imaginative, nonliteral ways; enactment of fictitious roles removed from everyday experiences (super heroes, fairy tales, cartoon characters).	Imagining; transforming materials and tasks; recognizing relations; making connections and associations; showing openness and flexibility; using imagery and symbolic representation.
Literacy (or may be procedural, depending on focus)	Books, story books with tapes, reading and writing activity.	Entering world of make-believe; expressing; practicing new roles.
Social orientation		
Dramatic play: Focus on relationships	Dolls and accessories, kitchen sets and other housekeeping props, dress-up clothes (may represent a variety of thematic roles).	Engaging in symbolic representation and verbal communication; enacting roles about life experiences; taking on adult roles/tasks, representing what is understood; practicing and testing understanding.
Dramatic play: Focus on social process	Miniature props: Cars, trains, bulldozers, and so on; buildings for community/occupational themes (houses, fire stations, airports, farm sets, mechanic garages).	Using props to enact roles and to imitate and recreate structures and procedures of society, exhibiting attention to social structure, roles, rules, and procedures.

Received September 16, 2005

Revision received January 29, 2008

Accepted February 26, 2008 ■



# Supporting Students' Motivation, Engagement, and Learning During an Uninteresting Activity

Hyungshim Jang  
University of Wisconsin—Milwaukee

The present study examined the capacity of 2 different theoretical models of motivation to explain why an externally provided rationale often supports students' motivation, engagement, and learning during relatively uninteresting learning activities. One hundred thirty-six undergraduates (108 women, 28 men) worked on an uninteresting 20-min lesson after either receiving or not receiving a rationale. Participants who received the rationale showed greater identified regulation, interest-enhancing strategies, behavioral engagement, and conceptual learning. Structural equation modeling was used to test 3 alternative explanatory models to understand why the rationale produced these benefits—an identified regulation model based on self-determination theory, an interest regulation model based on interest-enhancing strategies research, and an additive model that integrated both models. The data fit all 3 models; however, only the model that included rationale-enhanced identified regulation uniquely fostered students' engagement and hence their learning. Findings highlight the role that externally provided rationales can play in helping students generate the autonomous motivation they need to engage constructively in and learn from uninteresting, but personally important, lessons.

**Keywords:** rationale, autonomous motivation, identified regulation, interest regulation, self-determination theory

As students make the transition from elementary to middle school and from middle school to high school, their workload becomes greater, academic work increases in difficulty, grading becomes more stringent, and instruction becomes less personalized (Eccles & Midgley, 1989). Not surprisingly, students' academic motivation steadily declines following this transition, as children's mostly intrinsically motivated orientation gives way to adolescents' mostly extrinsically motivated orientation (Harter, 1981, 1982). In a similar vein, students—and especially older students—report finding the learning activities they encounter in school to be lacking in direct or personal relevance to their lives as well as unexciting, unappealing, overly complex and difficult, and/or more time consuming than they prefer (Anderman & Maehr, 1994; Eccles & Midgley, 1990; Eccles & Wigfield, 1995; Goodlad, 1984; Haladyna & Thomas, 1979; Hidi & Harackiewicz, 2000; Wigfield & Eccles, 2000). When students fail to value what they are asked to learn in school, the degree of student motivation to engage in the target learning activity significantly decreases (Legault, Green-Demers, & Pelletier, 2006; Murdock, 1999; Wigfield & Eccles, 2000). This devaluing process predicts students' subsequent minimal effort, poor concentration, indifference, and general withdrawal from the activity (Legault et al., 2006; Ntoumanis, Pensgaard, Martin, & Pipe, 2004; Vallerand, Fortier, & Guay, 1997; Vallerand et al., 1993). Consequently, teachers often find themselves facing a difficult motivational problem when they attempt to motivate students during uninteresting (from the stu-

dent's perspective), but potentially important (from the teachers' perspective), academic activities.

## The Relation Between Values and Academic Engagement and Performance

A substantial body of research on values and academic behaviors suggests that when students value a learning activity in terms of high task value, utility value, interest value, attainment value, instrumental value, future goals, future consequences, future time perspective, and/or intrinsic goals, they become increasingly likely to actively engage in that topic, to persist in that topic over time, to achieve highly, to show relatively sophisticated self-regulation, and to understand what they are trying to learn (DeVolder & Lens, 1982; Husman & Lens, 1999; Miller & Brickman, 2004; Miller, Greene, Montalvo, Ravindran, & Nichols, 1996; Shell & Husman, 2001; Vansteenkiste, Simons, Lens, Soenens, & Matos, 2005; Wigfield & Eccles, 2000). All these studies support the general conclusion that students tend to invest more effort and achieve more when a lesson is perceived to have personal importance or relevance.

The particular importance of these findings to teachers is that when teachers try to find ways to promote students' motivation during relatively uninteresting (but potentially important) learning activities, they can successfully do so by promoting task value. One way teachers can help students value the uninteresting, but important, learning task is by providing a rationale that (a) identifies the lesson's otherwise hidden value, (b) helps students understand why the lesson is genuinely worth their effort, (c) communicates why the lesson can be expected to be useful to them, and/or (d) helps students see or discover the personal meaning within a lesson. When successful, this instructional strategy can

Correspondence concerning this article should be addressed to Hyungshim Jang, Department of Educational Psychology, University of Wisconsin—Milwaukee, 709 Enderis Hall, Milwaukee, WI 53201. E-mail: hjang@uwm.edu

help create an opportunity for students to perceive, accept, and personally endorse—hence internalize into the self-system—the value of the learning activity.

### Purpose of the Present Study

Individual students bring their own influential characteristics (e.g., ability beliefs, personal interest, values) into the classroom. However, characteristics in the learning environment also affect students' motivational states, such as their interest, valuing, and effortful engagement. Focusing on these social-contextual conditions, the present study examined the effectiveness of two different theoretical models of motivation to explain why an externally provided rationale often supports students' motivation, engagement, and learning during relatively uninteresting learning activities. Because the predictions made in present study were derived from two theoretical models of motivation, the following sections review these two approaches to understanding when and why externally provided rationales can be expected to promote students' motivation during otherwise uninteresting lessons.

### The Identified Regulation Model

The identified regulation model, derived from self-determination theory (SDT; Deci & Ryan, 1985; Ryan & Deci, 2000) offers one explanation for why an externally provided rationale might facilitate motivation and engagement (see Deci, Eghrari, Patrick, & Leone, 1994; Reeve, Jang, Hardre, & Omura, 2002). SDT explains that when students find a learning activity to be important and personally meaningful to them—even if it is a relatively uninteresting thing to do—they experience a high-quality (i.e., autonomous) type of motivation referred to as “identified regulation” (Deci & Ryan, 1991; Ryan & Connell, 1989). Identified regulation refers to mostly internalized extrinsic motivation, as the individual has identified with the personal importance of an externally prescribed way of thinking or behaving and has thus accepted it as his or her own way of thinking or behaving (Deci & Ryan, 1991). Identified regulation is extrinsic because the activity is performed primarily because of its usefulness or instrumentality (work in order to develop a skill) rather than because it is interesting. It is self-determined because the student engages in the task willingly and for personal reasons rather than by being forced to engage the task because of external pressure. According to SDT's identified regulation model, the reason why an externally provided rationale promotes a student's internalization of task value into autonomous motivation of his or her own is because it provides the student with the information needed to self-identify with the activity as something the self willingly does because it is useful to the self (Deci et al., 1994; Reeve et al., 2002).

Deci et al. (1994) performed a laboratory experiment with a boring computer task (i.e., pressing the space bar on a keyboard whenever a light appeared on the computer screen) in which they manipulated the presence versus absence of three autonomy-supportive factors: a meaningful rationale, acknowledgment of the person's perspective (negative feelings participants might experience while undertaking such an unappealing task), and noncontrolling language that offered choice rather than pressure. The presence of a meaningful rationale did lead participants to perceive the task as an important one (relative to the condition with absence

of rationale). So, by itself, the rationale increased perceptions of task importance. However, when the rationale was communicated with controlling language and without an acknowledgement of the person's perspective, perceived autonomous motivation and extent of engagement were both low. In contrast, when the rationale was communicated with noncontrolling language and with an acknowledgment of negative feelings, autonomous motivation and engagement were both relatively high (Deci et al., 1994). Hence, for a rationale to promote engagement, it must promote not only high task importance but also perceptions of autonomous motivation (i.e., identified regulation).

In extending Deci et al.'s (1994) study, Reeve and his colleagues (2002) used a more academically authentic task to experimentally test whether the provision of a rationale when delivered in an autonomy-supportive way would increase students' effort during an uninteresting (but potentially important) learning activity (i.e., asking preservice teachers to learn conversational Chinese). In this study, identified regulation was conceptualized as a latent variable defined by the pair of indicators of perceived importance of the lesson and perceived autonomy while trying to learn it. In testing this mediation model, participants in the experimental condition were provided with a rationale as to why learning conversational Chinese might be a personally useful thing for them to do (i.e., gain a new teaching skill). Compared with participants not given this rationale, participants given the rationale showed greater effort. A motivational mediation model further showed that the reason why participants with the rationale showed the greater effort was because they felt an identified experience consisting of both a sense of ownership of the task (perceived autonomy) and a sense of task value (perceived importance). Thus, the reason why the externally provided rationale (delivered in an autonomy-supportive way) increased effort was because it allowed participants to experience higher identified regulation during the lesson.

### Interest Regulation Model

The interest regulation model, derived from Sansone and her colleagues' work (Sansone, Weir, Harpster, & Morgan, 1992), offers a second explanation for why an externally provided rationale might facilitate motivation and engagement. The interest regulation model explains that when people find a learning activity to be boring but inevitable, they generally attempt to regulate their interest by self-generating strategies designed to raise their immediate or situational interest to a level that is high enough to get through the otherwise uninteresting endeavor. These self-generated regulatory strategies are called interest-enhancing strategies (IESs; Sansone et al., 1992). Some of the most frequently used IESs include the strategies of setting a goal (Green-Demers, Pelletier, Stewart, & Gushue, 1998), varying the procedure so as to perform the same task in different ways (Sansone, Wiebe, & Morgan, 1999), working in the company of stimulating others such as friends (Isaac, Sansone, & Smith, 1999), and trying to make the task into a game (Wolters, 1998).

To test their hypothesis, Sansone et al. (1992) asked participants to perform a repetitive and boring activity—namely, repetitively copying pages of random letters. Prior to engaging in the boring copying task, participants were either provided or not provided



with a rationale (i.e., performing the task on a regular basis was said to yield health benefits). Sansone et al. predicted that participants performing the copying task with the knowledge of potential health benefits would be most likely to engage in the IESs because they had sufficient reason (the potential health benefit) to expend the effort. Sansone et al. found that hearing a rationale helped participants transform the otherwise boring task into a potentially more interesting one. For example, participants given the rationale performed the copying task more creatively (i.e., less repetitively). This modification, Sansone et al. argued, made the task temporarily more interesting. In a similar (but correlational) study, Wolters (1998) found that college students self-reported more frequent use of IESs to regulate their low motivation during uninteresting academic tasks. These students reported that when they had a strong need to study uninteresting lectures or textbook readings, they became more likely to “make studying into a game” or, simply, “try to make studying more interesting” (Wolters, 1998, p. 229).

Although not studied extensively, there is also evidence linking the use of IESs to greater subsequent effort and persistence. Sansone et al. (1992) found that students who generated an IES in their experimental study persisted longer at the repetitive letter-copying task than did students who did not use such a strategy. Wolters (1999), too, reported a positive correlation between high school students’ reported use of IESs and their degree of not only self-reported effort but also some specific study strategies, such as organization and monitoring.

In interpreting these findings, rationales produce motivational benefits because they prompt people to begin a mental search to find a way to make the uninteresting and unavoidable activity into something tolerable enough to get through it (Sansone & Smith, 2000). That is, when a task is uninteresting and when a rationale deems its performance to be a necessity, people generate more effort-promoting IESs.

## Hypotheses

As reviewed above, two independent explanations have been put forth to explain why externally provided rationales support students’ motivation and engagement. According to SDT, externally provided rationales promote engagement-enhancing identified regulation and internalization of task value into the self-system. According to the interest regulation model, externally provided rationales promote engagement-fostering IESs. Still, although these two different theoretical explanations are informative, it is not yet clear why students show increased engagement during uninteresting tasks in the presence of a rationale. To deepen psychological understanding, the two existing explanations need to be tested further—both independently and in combination with each other.

In the present study, it is hypothesized that, compared with participants not receiving an externally provided rationale, participants receiving an externally provided rationale will display a host of positive outcomes, including motivation, engagement, and conceptual learning. The dependent measures to index the quality of students’ learning experience are as follows. The first four measures assess the quality of students’ motivation and include concepts from both SDT (perceived autonomy, perceived importance) and the interest regulation model (IESs—essay and checklist).

IESs—essay represent the number of IESs participants spontaneously generated on an open-ended question, whereas IESs—checklist represent the number of IESs participants reported using from a prepared checklist of four possible strategies they might have used. The fifth and sixth measures assessed the extent of students’ engagement during the learning activity, as scored by trained raters (observers) during both the first (Time 1) and last (Time 2) 10 min of the learning session. The last two measures assessed the quality of participants’ learning, including measures of both factual and conceptual learning.

The reason why engagement was measured during two different intervals (Time 1 and Time 2) was to detect the enduring effect of the rationale on participants’ engagement over time. As shown in Sansone et al.’s (1992) and Reeve et al.’s (2002) studies, participants with the rationale (vs. its absence) were expected to maintain their task engagement longer compared with participants in the control group, either because IESs made the learning activity more tolerable (Sansone et al., 1992) or because participants identified with the task’s value (Reeve et al., 2002).

As to the two measures assessing the quality of students’ learning, factual learning refers to rote learning and the extent to which participants were able to repeat facts presented during the lesson whereas conceptual learning refers to understanding the core or main ideas discussed during the lesson. I predicted that although participants who receive a rationale will show greater conceptual learning than participants who do not receive a rationale, factual learning will not substantially differ between the two groups. This prediction is based on multiple tests of SDT that have shown that autonomous motivation (e.g., identified regulation) predicts conceptual, but not necessarily factual or rote, learning (Benware & Deci, 1984; Grolnick & Ryan, 1987; Vansteenkiste et al., 2005). This is so because students with autonomous motivation experience a deep, thoughtful, and task-orientated commitment toward learning, and these students process information in a more conceptual and integrative manner, compared with their counterparts. In contrast, these studies report that students with low autonomous motivation still show relatively high factual learning because factual or rote learning requires a more straightforward path to the solution, such as memorization (Benware & Deci, 1984; Grolnick & Ryan, 1987; Vansteenkiste et al., 2005).

To examine why a rationale supports students’ motivation, engagement, and learning during an uninteresting lesson, three possible explanatory models are proposed (see Figure 1). Model 1 depicts SDT’s identified regulation (or internalization) model. It is a motivational mediation model in which the rationale facilitates identified regulation, which, in turn, facilitates engagement, which, in turn, enhances conceptual learning. To operationally define identified regulation as a latent variable, the two indicators of perceived autonomy and perceived importance were assessed (following Reeve et al., 2002). In testing the identified regulation model (Model 1), the present study sought to replicate previous SDT research showing that rationales, when presented in autonomy-supportive ways, facilitate students’ identified regulation (internalization) and engagement. It is important to note, however, that the present study added a learning outcome to test whether the engagement engendered by the rationale and accompanying identified regulation would enhance students’ conceptual

Model 1: Identified Regulation Model



Model 2: Interest Regulation Model



Model 3: Additive Model

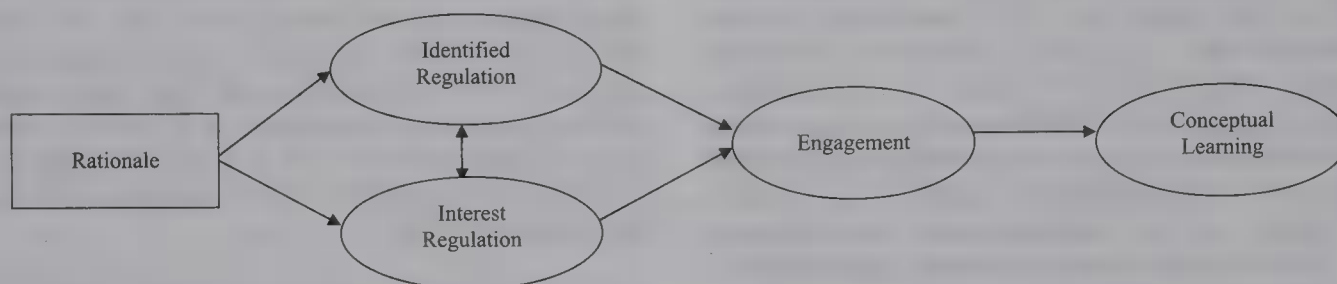


Figure 1. Three hypothesized models.

learning. Thus, this learning outcome was an important addition to Reeve et al.'s (2002) identified regulation model.

Model 2 depicts the interest regulation model. It also is a motivational mediation model. In this model, the rationale facilitates interest regulation through the creation and use of IESs, which, in turn, facilitate engagement, which, in turn, enhances conceptual learning. To operationally define interest regulation as a latent variable, I assessed two indicators of the creation and utilization of IESs. The first reported IESs using an open-ended essay format; the second reported the number of IESs checked off of a prepared list of four possible strategies. In testing the interest regulation model (Model 2), the present study sought to replicate previous interest regulation research showing that rationales prompt individuals to create and use IESs. Although the correlation between use of IESs and persistence or extended effort exists, no previous work in this area has experimentally tested the relation that interest regulation has to a learning outcome. Thus, this learning outcome was an important addition to the interest regulation model.

Model 3 depicts an additive model in which the rationale facilitates both identified regulation and interest regulation, both of which, in turn, contribute a unique positive effect on engagement. Further, engagement enhances conceptual learning. This additive model simply combines Model 1 (identified regulation model) and Model 2 (interest regulation model) but, in doing so, allows the key motivational construct from each model to compete to explain the variance in engagement. This will allow a determination of whether identified regulation and interest regulation are additive or whether only one of these motivational processes is sufficient to explain engagement.

## Method

### Participants

Participants included 136 college students (108 women, 28 men) recruited from sections of an introductory educational psychology class at a large Midwestern university. All participants were enrolled in the teacher certification program and were preparing to become teachers after graduation. In exchange for their participation, each participant received extra course credit.

### Uninteresting Activity

The learning activity was selected on the basis of three characteristics: (a) Students would not generally perceive it to be intrinsically interesting, (b) it represented an ecologically valid and school-like lesson, and (c) it possessed hidden value and relevance so that participants could potentially find some personal utility within it. For the uninteresting—yet potentially worthwhile—ecologically valid lesson, I imported a lesson on correlations from an introductory-level statistics textbook for undergraduates (Frederick & Wallnau, 2002). The learning material featured a six-page text-based lesson that covered the following topics: correlation coefficient, scatterplots, correlation and prediction, and correlation and causation. Learning about correlations can be an interesting activity, but for the purposes of the present study, the lesson was presented in an uninteresting format in that it was designed to be both monotonous (following Berlyne, 1966) and void of interest-enhancing embellishments (following Parker & Lepper, 1992). Pilot testing with 32 participants indicated that participants found the lesson to be relatively uninteresting ( $M = 2.17$  on a 7-point



Likert scale with 1 = *not at all interesting* and 7 = *extremely interesting*). The pilot test also confirmed that these 32 participants found 20 min to be an appropriate amount of time to learn the lesson ( $M = 6.25$  on a 7-point Likert scale with 1 = *not at all, too little time* and 7 = *very appropriate*).

### Rationale

An externally provided rationale is a verbal explanation as to why putting forth effort during an activity is a useful and worthwhile thing to do. Past motivation research makes it clear, however, that some rationales are better than others when it comes to engaging students in learning activities. Rationales that are presented in controlling ways ("Do it because I said so" or "Do it because there will be a test") do not help students internalize the value of the activity (Deci et al., 1994; Reeve et al., 2002). In contrast, rationales that are provided in an autonomy-supportive way do help students internalize the value of the activity (Deci et al., 1994; Reeve et al., 2002). For this reason, the rationale manipulation was operationally defined as an externally provided rationale delivered in an autonomy-supportive way. To deliver the rationale in an autonomy-supportive way, I communicated the rationale to participants with both noncontrolling language and the acknowledgement of possible negative feelings.

The content of the rationale used in the present study—namely, that learning correlations is an opportunity to cultivate useful teaching knowledge—was constructed on the basis of rationales offered in several different introductory statistics textbooks explaining the merits of learning about correlations. Once constructed, the rationale was pilot tested with a different group of 35 participants by confirming that they perceived it to be relatively convincing ( $M = 5.78$  on a 7-point Likert scale with 1 = *not at all convincing* and 7 = *very convincing*). The rationale (with embedded noncontrolling language and the acknowledgement of negative feelings) was as follows:

Learning about correlations has been shown to be useful. Those participants who have learned about correlations featured in today's lesson have reported that it helped them become a more reflective teacher. They became more reflective because the lesson learned helps them to see relationships that the naked eye might miss. Correlations summarize vast information to help teachers explore issues, answer questions, solve problems, and make decisions. This is why many educational journals, internet educational websites, and textbooks present information using correlations. For example, a quick glance at the newspaper yields statistics that communicate research findings on new teaching methods, parent-school relations, new instructional technologies, student-teacher relationships, teachers' average income, and so on.

Learning about correlations may not be much fun for some of you. So it is understandable that you might not find it very interesting. Nonetheless, today's lesson is particularly designed to help you think about how two variables might or might not be related to one another or whether one variable predicts or causes the other. Once learned, the correlations featured in today's lesson will open the door for you to gain useful skills, ones that will be very handy when you need to interpret information presented through statistical tools. This is the reason why you are being asked to concentrate, persevere, and try hard on the lesson.

### Procedure

All materials, including the rationale, were presented in booklets. The reason the material was presented in booklets (instead of by an experimenter) was to control for extraneous factors, minimize potential demand characteristics, and increase the study's internal validity (following Deci et al.'s, 1994, research paradigm). Participants were randomly assigned to one of two groups (control condition, experimental condition) and were tested in small groups, with an average of 6 individuals per group. Participants were seated such that they had no contact with other participants. Before the experiment began, all participants completed a consent form and preexperimental questionnaire assessing demographic information.

During the experimental session, two main events occurred. First, the experimenter announced a 3-min introductory period. During this period, the independent variable was manipulated in that each participant received a sheet containing one of two instructions (one with instructions and a rationale, one with instructions and no rationale) according to a random sequence within session. Participants were asked to read the instruction silently. The sheet began with the following general instruction (following Benware & Deci, 1984):

Please use these materials (that you will be given) to learn about correlations. The following topics are covered in the lesson: correlation, correlation coefficient, scatterplots, correlation and prediction, and correlation and causation. Read and study the text materials in the same manner that you would read and study any text assigned in one of your college courses. Please feel free to write on the material or to take notes on the papers provided.

Participants in the experimental condition received additional instruction that included the rationale. Participants in the control condition did not receive this additional instruction.

Second, a 20-min individual learning session followed. During this learning session, all participants received the same six-page booklet of learning materials and were asked to study the materials for 20 min. Participants read and studied silently while working independently and at their own pace. During this period, two trained raters who were naive to the experimental condition independently and surreptitiously scored how behaviorally engaged versus disengaged each participant appeared to be. The raters sat nonintrusively in the corner of the classroom and made independent ratings. The raters scored participants' engagement objectively so that the present study could advance prior studies in this area of research that had used only self-reported engagement scores. After the learning period ended, the experimenter announced that the study time had ended and administered the postexperimental questionnaire. After the postexperimental questionnaires were collected, the experimenter administered an unannounced test to assess learning. Last, all participants received a debriefing.

### Measures

*Identified regulation.* Identified regulation consists of the compound psychological experience of high perceived importance and high perceived autonomy (Reeve et al., 2002). Perceived autonomy was assessed with the nine-item Perceived Autonomy



Scale (Reeve, Nix, & Ham, 2003). The scale has three items to assess an internal perceived locus of causality (e.g., "During the lesson, I felt I was doing what I wanted to be doing"), three items to assess volition (e.g., "During the lesson, I felt free"), and three items to assess perceived choice over one's actions (e.g., "I felt I had control to decide what to do and whether to do it"). Each item used a 7-point Likert scale (1 = *not at all true*, 7 = *very much true*), and scores from these nine items intercorrelated highly enough to average into a single score ( $\alpha = .83$ ). Scores on the Perceived Autonomy Scale have been shown to be valid in that they are sensitive to autonomy-supportive versus controlling teaching styles and predict various measures (self-report, behavioral) of intrinsic motivation (Reeve et al., 2003).

Perceived importance was assessed with a perceived importance scale from a previous study (Reeve et al., 2002). Each of the four items used a 7-point Likert scale (1 = *not at all true*, 7 = *very much true*) with the stem, "Learning the lesson about correlations was . . . : an important thing to do, pointless—a waste of my time (reverse scored), valuable, and worthwhile—it was time well spent. Scores from these four items intercorrelated highly enough to average into a single score ( $\alpha = .87$ ). Scores on this perceived importance scale have been shown to be valid in that they are sensitive to teacher-provided rationales and able to predict engagement during a learning activity (Reeve et al., 2002).

**Interest regulation.** Interest regulation consists of the two indicators of the number of IESs participants reported using during the 20-min study session. The IESs—essay measure used the open-ended question,

During the 20 minute lesson, what did you do to make this learning activity a more interesting thing to do (if anything)? Perhaps you didn't do anything to make the lesson seem more interesting, but if you did use such a strategy, please write your strategy or strategies in the space below.

Two trained independent raters coded the responses using a scoring system derived from empirical work on IESs (following Sansone et al., 1992). From these essays, raters identified and scored the four nominated IESs: (a) set a goal (e.g., "finish prior to a time limit"), (b) used a fantasy context (e.g., "pretended to teach the lesson to someone else"), (c) introduced variety within the task (e.g., "varied how I did the task"), and (d) added stimulation (e.g., "drew a diagram"). Interrater reliabilities were high ( $r = .92$ ), so the raters' scores were averaged into a single score, called IESs—essay.

From pilot testing and from Sansone and colleagues' (1992) work, which IESs participants might use could be anticipated. Hence, a checklist was prepared to score participants' responses on a 4-point scale. The first item asked the participant to check whether or not he or she set a goal: "I set a goal for myself." The second item asked whether the participant used a fantasy context: "I imagined, pretended, or fantasized myself teaching or explaining this material to someone (e.g., my students)." The third item asked whether the participant introduced variety within the task: "I varied the task in some way (e.g., by switching my attention from one part of the text to another part)." Last, the fourth item asked whether the participant added stimulation: "I drew a picture or diagram to stimulate or entertain myself." Participants were asked to check off which specific strategies they used among the four listed. Checking any one individual strategy constituted a separate

point on the scale. For instance, a student who reported using a fantasy context but not goal setting, variety, or stimulation scored a 1 on the IESs—checklist measure. The checklist appeared on the postexperimental questionnaire after participants had already completed the open-ended essay question.

To validate both measures, participants completed a three-item self-report measure of interest (Williams, Wiener, Markakis, Reeve, & Deci, 1994). Each item used the same 7-point response scale (1 = *not at all true*, 7 = *very much true*) with the stem, "Please rate 'learning about correlations' as an activity:" it held my full and constant attention, it stimulated my curiosity without interruption, and it was very interesting. These three items were averaged into a single score for interest ( $\alpha = .90$ ). This measure has been shown to be valid in that scores predict behavioral measures of intrinsic motivation (Reeve, 1989) and career choice decisions (Williams et al., 1994). In the present study, this self-report interest measure correlated significantly with the use of IESs as reported on both the essay,  $r(136) = .34, p < .01$ , and checklist,  $r(136) = .26, p < .01$ , measures. These positive correlations are important because they confirm that the degree to which these strategies were used was positively associated with interest.

**Engagement.** Raters scored three aspects of participants' engagement during the lesson (based on Skinner & Belmont, 1993): on-task attention, effort, and persistence. To score these three behavioral expressions of engagement, two trained raters used a rating sheet with 7-point bipolar scales. For on-task attention, the bipolar descriptors were dispersed—off task (scored as 1) versus focused—on task (scored as 7). For effort, the bipolar descriptors were passive, slow, or minimal effort (scored as 1) versus active, quick, or intense effort (scored as 7). For persistence, the bipolar descriptors were gives up easily during challenge, failure, or confusion (scored as 1) versus persistent (scored as 7). During the 20-min lesson, two trained raters who were naive to the experimental condition independently made two separate ratings—one during the first 10 min and the second during the last 10 min. The reason why engagement was measured during two different times was, as mentioned earlier, to detect the effect of the rationale on engagement over time.

For both rating periods, raters' scores correlated highly with one another. The intercorrelations between the two raters at Time 1 (first 10 min) were as follows: attention,  $r = .76, p < .01$ ; effort,  $r = .75, p < .01$ ; and persistence,  $r = .76, p < .01$ . The intercorrelations between the two raters at Time 2 (last 10 min) were as follows: attention,  $r = .90, p < .01$ ; effort,  $r = .88, p < .01$ ; and persistence,  $r = .86, p < .01$ . Because interrater reliabilities were high, the pairs of scores from the two raters were averaged to form a single score for attention, effort, and persistence at each time period. Once done, the three engagement ratings (attention, effort, and persistence) were averaged into the two following engagement scores: engagement at Time 1 (three items,  $\alpha = .96$ ) and engagement at Time 2 (three items,  $\alpha = .97$ ).

**Learning.** Following Benware and Deci (1984), the dependent measure to assess the two types of learning was scored from a 14-item multiple-choice examination. Each question was designed to measure either factual learning or conceptual learning of the material. To construct this measure, experts in the statistics department at a major university read the text and created 14 factual questions to assess recognition of facts and 14 questions to assess conceptual understanding of the information. Two expert raters



then independently scored each question categorically as either a factual question or a conceptual question. Only those items that received an identical factual–conceptual classification by the two experts were used in the study. The final test featured seven factual questions and seven conceptual questions. An example of a factual learning question was, “A correlation coefficient indicates the . . .” (followed by four response options). An example of a conceptual learning question was, “Which of the following questions is best suited for correlational research?” (followed by four response options). Possible scores for both the factual learning test and the conceptual learning test could range from 0 to 7.

## Results

Descriptive statistics for the eight dependent measures appear in Table 1, broken down by experimental condition. A series of *t* tests was used to test for mean differences between the experimental and control groups. To conduct seven independent tests (eight dependent measures minus factual learning) and still protect against making a Type 1 error, I calculated what each testwise alpha level must be to produce an overall experimentwise (exp) alpha level of .05. With seven tests, the  $\alpha_{\text{exp}}$  was inflated to .30 (using  $\alpha_{\text{test}} = .05$ , one-tailed; based on Hays’s, 1984, formula of  $\alpha_{\text{exp}} = 1 - [1 - .05]^7$ ). So, to readjust the inflated .30  $\alpha_{\text{exp}}$  back down to .05, I computed what each  $\alpha_{\text{test}}$  needed to be for each correlation test. This value was .014 (based on Hays’s, 1984, formula of  $\alpha_{\text{test}}/\text{number of tests}$ , or  $.05_{\text{one-tailed}}/7$ ). Because my hypotheses were directional, using a one-tailed test was both more suitable and more powerful (Hopkins, Glass, & Hopkins, 1987; Minium, King, & Bear, 1993).<sup>1</sup> Prior to conducting these *t* tests, I first explored whether gender influenced any motivation, engagement, or learning measure. A series of *t* tests were performed, and results showed that gender did not influence any of the study’s eight dependent measures. The data for each dependent measure were therefore collapsed across gender.

The provision of an externally provided rationale enhanced motivation, and this was true for all four measures, including perceived autonomy ( $d = 0.55$ ), perceived importance ( $d = 0.71$ ), IESs–essay ( $d = 0.56$ ), and IESs–checklist ( $d = 0.42$ ). The provision of an externally provided rationale enhanced engagement, and this was true for both engagement at Time 1 ( $d = 0.44$ ) and engagement at Time 2 ( $d = 0.64$ ). For learning, the provision of an externally provided rationale enhanced conceptual learning ( $d = 0.39$ ) but not factual learning.

### *Effect of the Rationale on Engagement Over Time*

As reported, raters scored the participants who received the rationale as significantly more engaged during the 20-min lesson than participants who did not receive the rationale. To test whether the rationale supported an engagement-fostering benefit over time, I performed a repeated-measures analysis using the presence versus absence of the rationale as the between factor and the rating period (Time 1, Time 2) as the within factor. Both the rationale,  $F(1, 134) = 13.08$ ,  $p < .001$ , and the rating period,  $F(1, 134) = 140.63$ ,  $p < .001$ , were individually significant, as the rationale facilitated engagement and engagement decreased over time. Of more importance, however, the Rationale  $\times$  Rating Period interaction effect was significant,  $F(1, 134) = 7.72$ ,  $p < .01$ , as the rate

of disengagement from the uninteresting lesson for participants in the control group was significantly more pronounced than it was for participants who received the rationale, as shown in Figure 2. This is an important finding because it indicates that the motivational benefits from the rationale (i.e., identified regulation, interest regulation) became increasingly important to sustaining engagement as the lesson continued over time.

### *Test of Hypothesized Models*

The three hypothesized models were tested with structural equation modeling using LISREL 8.51 (Jöreskog & Sörbom, 2001). To test the fit of the data to the hypothesized models, I followed the two-step approach recommended by Anderson and Gerbing (1998). First, to determine whether the indicators related satisfactorily to the latent variables, I performed a confirmatory factor analysis to assess the fit of the measurement model. Second, the series of three hypothesized models (see Figure 1) were tested as structural models. To evaluate the fits of the measurement and structural models, I relied on a set of five test statistics. Traditionally, a nonsignificant chi-square serves as the basic test of whether a model adequately describes the data (Bollen & Long, 1993); however, I further included a set of fit indices because they often provide a better evaluation of model fit than does the chi-square statistic (Bentler & Bonett, 1980; Marsh, Balla, & McDonald, 1988). Those four fit indices were the root-mean-square error of approximation (RMSEA), the root-mean-square residual (RMR), the nonnormed fit index (NNFI), and the comparative fit index (CFI). RMSEA and RMR are summary statistics for the residuals, so the lower the number, the better (i.e.,  $\text{RMR and RMSEA} < .05$ , down to a possible low of 0; Hu & Bentler, 1999). NNFI and CFI compare the lack of fit of the theoretical model to the independence model, so the higher the number, the better (i.e.,  $\text{NNFI and CFI} > .95$ , up to a possible high of 1; Hu & Bentler, 1999).

According to the chi-square statistic and the goodness-of-fit indices, the measurement model fit the observed data well,  $\chi^2(9, N = 136) = 5.58$ , *ns*,  $\text{RMSEA} = .00$ ,  $\text{RMR} = .03$ ,  $\text{NNFI} = 1.00$ ,  $\text{CFI} = 1.00$ . In examining the parameter estimates, each measure–indicator loaded significantly and positively on its appropriate latent factor.

To conduct the main structural model analyses, I categorically scored the provision of a rationale—the manipulated predictor variable—as 0 for absence of a rationale and as 1 for the provision of a rationale. Means, standard deviations, and intercorrelations among the eight measures included in the hypothesized models appear in Table 2.

**Hypothesized Model 1.** The identified regulation model proposed that the rationale would enhance participants’ identified regulation, which would increase participants’ engagement, which, in turn, would enhance their learning (see Figure 3). This model fit the

<sup>1</sup> One-tailed tests were used because participants’ initial orientation to the lesson in terms of motivation and prior knowledge in both groups (experimental vs. control) was already low. Hence, it was doubtful that participants who received a rationale in the present study would show a significant decrease in these items compared with participants in the control group. Also, no previous study has shown that participants with a rationale scored lower on a measure of motivation, engagement, or learning than did participants without a rationale (Benware & Deci, 1984; Deci et al., 1994; Reeve et al., 2002; Sansone et al., 1992).

Table 1  
Means (and Standard Deviations) for Each Dependent Measure by Experimental Condition

Dependent measure	Possible range	Experimental condition		<i>t</i> (134)
		Rationale absent ( <i>n</i> = 67)	Rationale present ( <i>n</i> = 69)	
Perceived autonomy	1-7	4.28 (1.07)	4.87 (1.07)	3.57*
Perceived importance	1-7	3.89 (1.22)	4.66 (0.94)	4.30*
Interest-enhancing strategies—Essay	0-4	0.38 (0.52)	0.69 (0.58)	3.26*
Interest-enhancing strategies—Checklist	0-4	0.99 (0.86)	1.35 (0.84)	2.49*
Behavioral engagement, Time 1	1-7	5.44 (1.10)	5.86 (0.81)	2.50*
Behavioral engagement, Time 2	1-7	3.70 (1.66)	4.72 (1.51)	3.73*
Factual learning	1-7	5.70 (1.27)	6.07 (1.18)	1.76
Conceptual learning	1-7	5.27 (1.45)	5.87 (1.14)	2.69*

\*  $p < .014$ .

observed data well,  $\chi^2(7, N = 136) = 2.65$ , *ns*, RMSEA = .00, RMR = .02, NNFI = 1.00, CFI = 1.00. As shown in Figure 3, each of the hypothesized paths within the identified regulation model was significant and in the predicted direction, as the rationale predicted identified regulation ( $\beta = .42$ ,  $p < .01$ ), and identified regulation predicted engagement ( $\beta = .33$ ,  $p < .01$ ), which, in turn, predicted learning ( $\beta = .45$ ,  $p < .01$ ). Further, the direct (unmediated) path from the rationale to engagement was not significant ( $\beta = .19$ , *ns*), showing that identified regulation, rather than the provision of the rationale per se, best explained extent of engagement. The overall identified regulation model explained 19% of the variance in engagement and 20% of the variance in learning.

**Hypothesized Model 2.** The interest regulation model proposed that the rationale would enhance participants' interest regulation, which would increase their engagement, which, in turn, would enhance their learning (see Figure 4). This model fit the observed data well,  $\chi^2(7, N = 136) = 6.54$ , *ns*, RMSEA = .00, RMR = .04, NNFI = 1.00, CFI = 1.00. As shown in Figure 4, each of the hypothesized paths within the interest regulation model was significant and in the predicted direction, as the rationale predicted interest regulation ( $\beta = .41$ ,  $p < .01$ ), and interest regulation predicted engagement ( $\beta = .25$ ,  $p < .01$ ), which, in turn, predicted learning ( $\beta = .44$ ,  $p < .01$ ). However, unexpectedly, the direct path from the rationale to engagement remained significant ( $\beta = .22$ ,  $p < .05$ ),

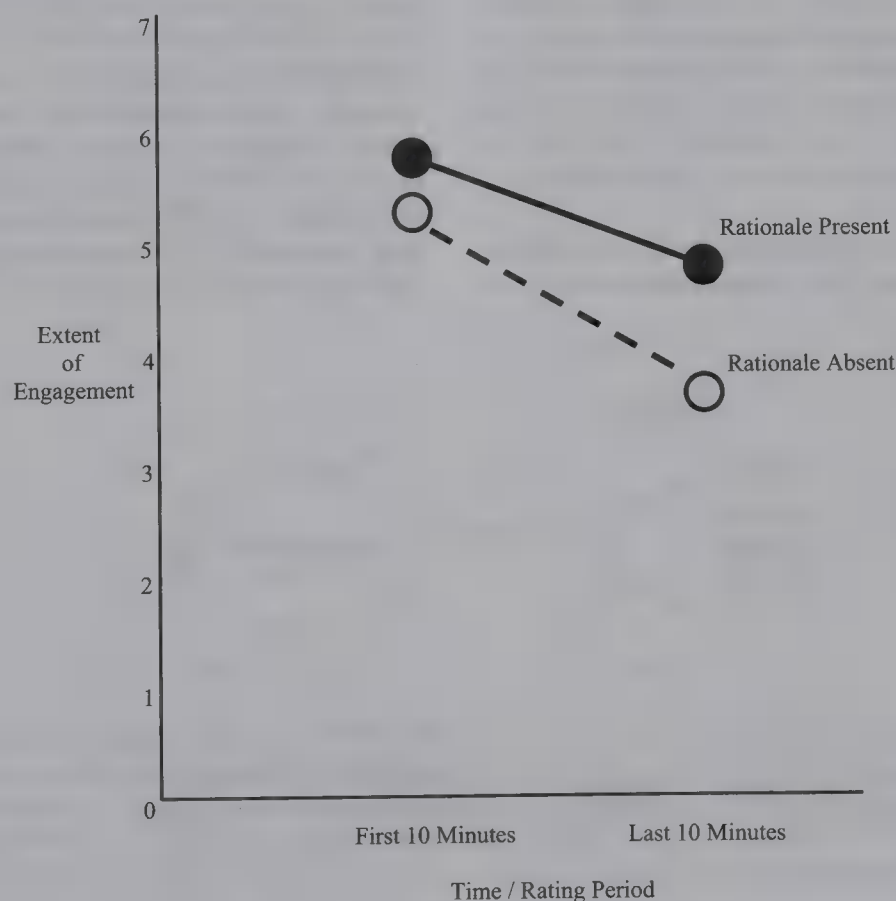


Figure 2. Effects of rationale and rating period on participants' engagement.



Table 2  
Descriptive Statistics and Intercorrelation Matrix for all Dependent Measures Included in Hypothesized Models

Dependent measure	M	(SD)	1	2	3	4	5	6	7	8
1. Rationale	0.51	(0.50)	—							
2. Perceived autonomy	4.57	(1.00)		—						
3. Perceived importance	4.28	(1.10)			—					
4. Interest-enhancing strategies—Essay	0.54	(0.57)				—				
5. Interest-enhancing strategies—Checklist	1.17	(0.87)					—			
6. Behavioral engagement, Time 1	5.65	(0.98)						—		
7. Behavioral engagement, Time 2	4.42	(1.66)							—	
8. Conceptual learning	5.57	(1.33)								—

Note. The possible range for each dependent measure was from 1 to 7, except for Measures 4 and 5; the possible range for these measures was from 0 to 4. *N* = 136. \* *p* < .05, one-tailed. \*\* *p* < .01, one-tailed.

showing that interest regulation only partially mediated the effect that the provision of a rationale had on engagement. The overall identified regulation model explained 16% of the variance in engagement and 20% of the variance in learning.

**Hypothesized Model 3.** The additive model proposed that the rationale would enhance both participants' identified regulation and interest regulation (see Figure 5). This model further predicted that both identified regulation and interest regulation would each contribute uniquely and positively to predicting participants' engagement. Finally, the additive model predicted that extent of engagement would predict learning. The additive model fit the observed data well,  $\chi^2(15, N = 136) = 9.05, ns$ , RMSEA = .00, RMR = .03, NNFI = 1.00, CFI = 1.00. As shown in Figure 5, the rationale significantly predicted both identified regulation ( $\beta = .41, p < .01$ ) and interest regulation ( $\beta = .41, p < .01$ ). In the prediction of engagement, however, identified regulation predicted engagement ( $\beta = .32, p < .01$ ) but interest regulation did not ( $\beta = .02, ns$ ). Further, the direct (unmediated) path from the rationale to engagement was not significant ( $\beta = .18, ns$ ). The overall additive model explained 19% of the variance in engagement and 20% of the variance in learning. Of note, the key motivational constructs were not additive in the effects on engagement; when they competed for variance, only identified regulation predicted engagement.

**Conclusion.** All three structural models fit the data equally well, all three models accounted for a comparable amount of the

variance in both engagement and learning, and no model fit the data significantly better than did another. The reason why Model 2 is the less favored model is because it needed the rationale to explain engagement, whereas Models 1 and 3 did not. As shown in Model 3, the reason why rationale predicted engagement in Model 2 was because the rationale facilitated the identified regulation process. Hence, although all three models fit the data comparatively well, the pattern of significant and nonsignificant paths to engagement made it clear that engagement was facilitated by identified regulation and not by interest regulation.

Discussion

Recognizing that an externally provided rationale can promote students' motivation and engagement during an uninteresting lesson, the present study sought to provide a theoretical and comprehensive understanding of the functional motivational significance that an externally provided rationale can have on students' engagement and learning. Findings showed that an externally provided rationale, when delivered in an autonomy-supportive way, promoted a relatively high-quality learning experience for participants, as assessed by their motivation, engagement, and conceptual learning (see Table 1). These findings confirm past research findings showing the motivational benefits of externally provided rationales during uninteresting learning activities. Findings in the

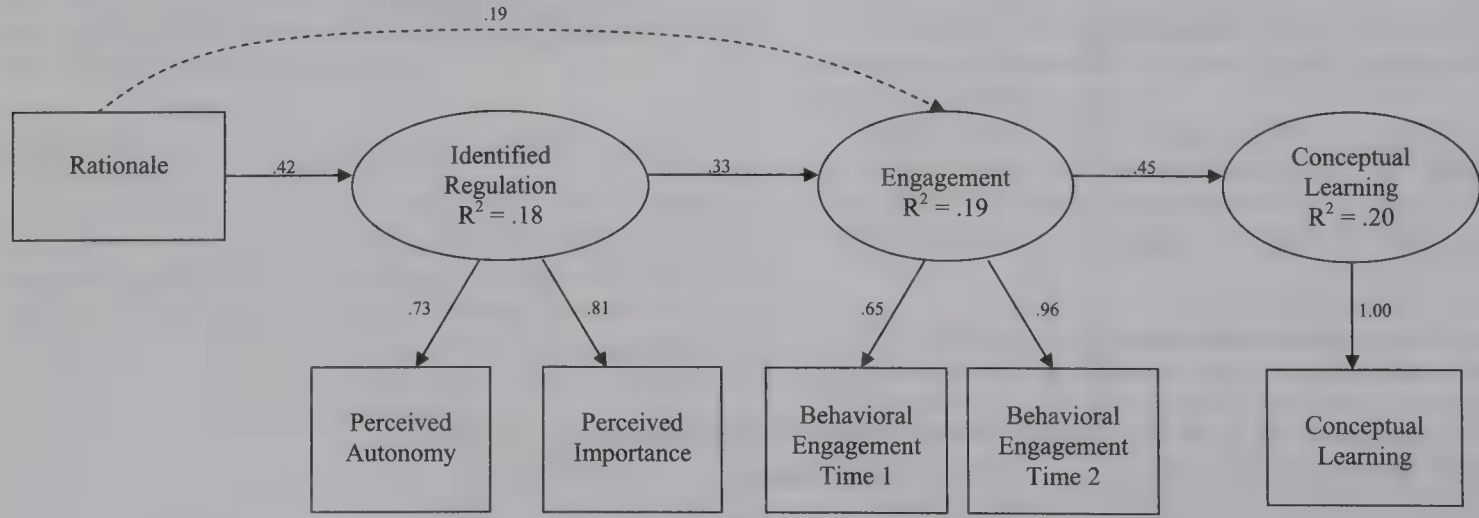


Figure 3. Standardized parameter estimates from the LISREL analysis of Model 1: the identified regulation model. Solid lines represent significant paths, *p* < .05; dashed lines represent nonsignificant paths.

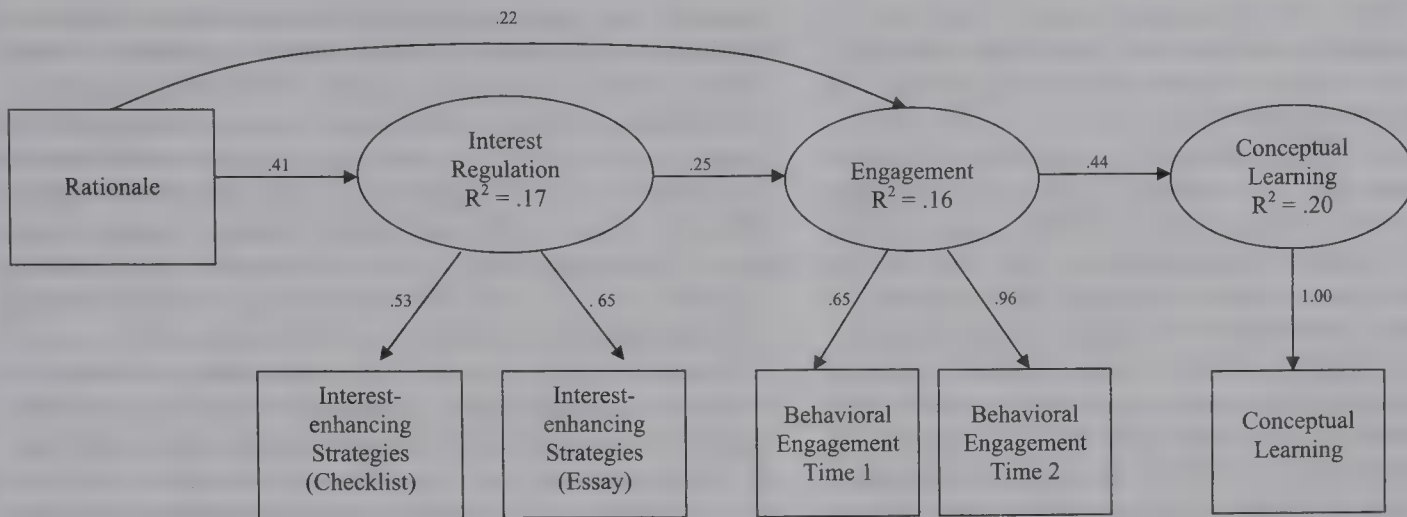


Figure 4. Standardized parameter estimates from the LISREL analysis of Model 2: the interest regulation model. Solid lines represent significant paths,  $p < .05$ ; dashed lines represent nonsignificant paths.

present study further allow psychologists to answer the question of why such rationales generate these benefits.

#### *Why a Rationale Supports Students' Motivation, Engagement, and Learning*

The empirical test of the identified regulation model (see Figure 3) showed that the identified regulation model fit the data well. Accord-

ing to SDT, rationales facilitate engagement and learning because a rationale, when communicated in an autonomy-supportive way, reveals an activity's value and personal benefit (Ryan & Deci, 2000, 2002). Such personal relevance information helps participants identify with and internalize the value of the task (identified regulation), and this internalization allows participants to engage volitionally in the learning activity. Thus, although the activity itself was inherently uninteresting, the externally provided rationale facilitated

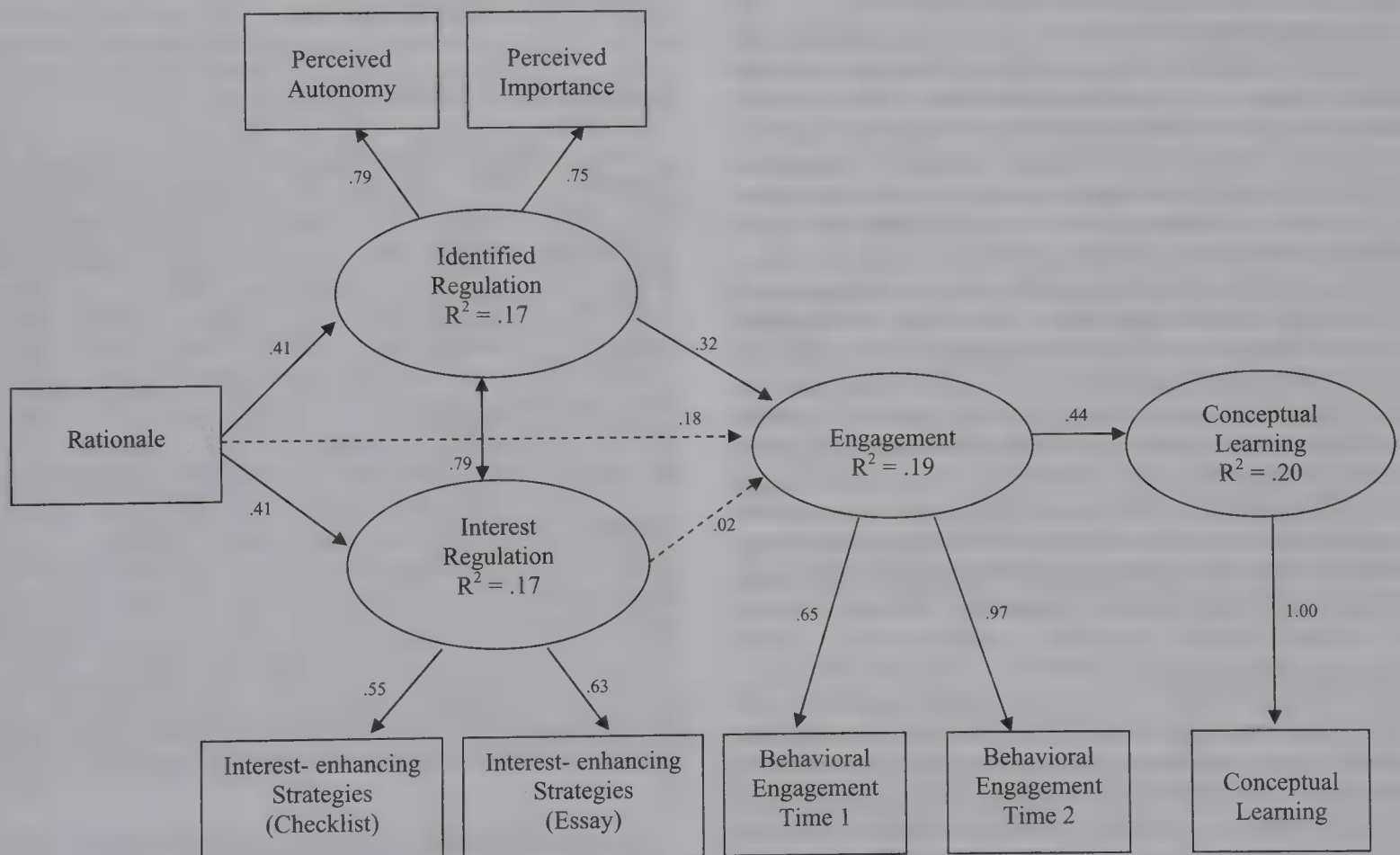


Figure 5. Standardized parameter estimates from the LISREL analysis of Model 3: the additive model. Solid lines represent significant paths,  $p < .05$ ; dashed lines represent nonsignificant paths.



participants' capacity to take on the externally provided rationale as their own self-endorsed reason to try hard. Once experienced, identified regulation was largely an engagement-fostering process, and extent of engagement explained learning.

The empirical test of the interest regulation model (see Figure 4) also fit the observed data well. According to the interest regulation model (Sansone et al., 1992), rationales facilitate engagement and learning because a rationale helps participants see the otherwise uninteresting activity as a necessary undertaking. Perceiving an uninteresting activity as a necessity prompts people to generate the IESs they need to make the activity tolerable. Although interest regulation did promote engagement, the rationale unexpectedly exerted a direct effect on engagement as well. This means that interest regulation explained only part of the reason why the rationale promoted engagement. Examination of the additive model revealed the reason why the rationale continued to exert a direct effect on engagement in the interest regulation model—namely, because this model did not include the important motivational mediating variable of identified regulation.

The empirical test of the additive model (see Figure 5) showed that the additive model fit the data well. As expected, the provision of a rationale promoted both identified regulation and interest regulation. When both these effects were considered together, the identified regulation effect on engagement was significant whereas the interest regulation effect was not. That is, adding the interest regulation path to engagement did not allow the additive model to explain any additional variance in engagement (or learning) beyond that explained by the identified regulation model itself. Hence, the additive model simply restates the identified regulation model but adds the tangential path from the rationale to enhanced interest regulation.

The findings reported for the additive model suggest that the reason why interest regulation has predicted task persistence in previous studies (Sansone et al., 1999; Werner & Makela, 1998) is probably because of the close association IESs have with identified regulation ( $\beta = .79, p < .01$ ; see Figure 5). Hence, although IESs help make an otherwise uninteresting learning experience more tolerable (less boring), they do not necessarily enhance engagement above and beyond the engagement-fostering properties of identified regulation. This is not to say that interest regulation does not play a meaningful role in the experience of uninteresting lessons. For instance, the use of IESs was significantly correlated with relatively high interest. Such a gain in positive emotion and subjective well-being is important and meaningful for its own sake. However, in the present study, interest regulation did not serve as an independent (unique) engagement-fostering strategy. This interpretation is consistent with Burton, Lydon, D'Alessandro, and Koestner's (2006) finding that identified regulation uniquely predicts effort and achievement (but interest regulation does not), whereas interest regulation uniquely predicts psychological well-being (but identified regulation does not).

### *Classroom Implications*

To facilitate students' motivation, rationales need to produce two effects: Students need to see the importance and personal utility within the task, and students need to perceive high autonomy while working on that task. The content of an externally provided rationale accomplishes the first purpose, whereas the way it is communicated—in an autonomy-supportive way—accomplishes the second purpose. When rationales are communicated in an autonomy-

supportive way, students are more likely to perceive the utility message within the rationale as a conduit for autonomy support. That is, students are likely to view the purpose—the functional significance (Deci & Ryan, 1991)—of the rationale as an external contingency intended to support their autonomy. If the same rationale were not delivered in an autonomy-supportive way, then it would not be expected to facilitate autonomy and hence identified regulation. The present study did not test this later assertion, but previous work by Deci and his colleagues (1994) shows that when rationales are presented in controlling ways, they fail to engender engagement-fostering benefits.

The provision of an externally provided rationale gives teachers an instructional strategy capable of fostering autonomous motivation, engagement, and subsequent learning during those lessons that teachers expect students might find relatively uninteresting. Being an external contingency, the rationale promotes extrinsic motivation but does so in a way that supports congruence between students' subjective feelings and their behavior directed toward the task. Most extrinsically motivating instructional strategies, such as extrinsic rewards, typically induce students into a compliance mode that places their subjective feelings ("this is boring") at odds with their engagement behavior (Joussemet, Koestner, Lekes, & Landry, 2005). Extrinsically motivating instructional strategies enhance engagement best when they allow harmony or congruence between students' inner motivational resources ("I want to do this") and their task engagement (spending 20 min studying the lesson). The externally provided rationale used in the present study worked as an effective extrinsically motivating strategy because it allowed students' experience of autonomous motivation (perceived importance, perceived autonomy) to be the motivational foundation that determined the extent of their task engagement and subsequent learning.

### *Implications for Subjective Task Values*

Value researchers emphasize three major contributors to (or components within) subjective task value: (a) extrinsic utility value, which is a task's perceived usefulness in accomplishing some desired end state, such as a career goal; (b) interest value, which is the task's perceived capacity to generate a sense of enjoyment; and (c) attainment value, which is a task's sense of importance to the individual's underlying self-system (Eccles & Wigfield, 1995).<sup>2</sup> Because this research has relied almost exclusively on nonexperimental survey-based research, the teacher's potential role in enhancing students' valuing of classroom activities has been largely unexplored (as researchers typically examine how privately held values correlate with task choices, such as whether or not to take a math class; Meece, Wigfield, & Eccles, 1990). The present findings therefore can potentially offer some unique insights into the process in which educators attempt to transfer objective task value (what the school values) to subjective task value (what the student values; Eccles & Wigfield, 1995).

Translating these three contributors to subjective task value into the concepts featured in the present experimental research aligns extrinsic utility value with the contents of the externally provided

<sup>2</sup> In addition to these three contributors, additional components, such as costs, subject matter appreciation, and/or future goals, are also possible contributors to subjective task value (Brophy, 1999; Husman & Lens, 1999; Miller & Brickman, 2004; Wigfield & Eccles, 1992).



rationale, interest value with interest regulation, and attainment value with identified regulation. To the extent that these concepts are interchangeable, three insights emerge in regard to the educational effort to promote students' subjective task value during uninteresting academic lessons.

First, the effort to directly communicate a task's extrinsic utility value to students can be expected to fail more often than not. This skepticism stems from previous rationale-based research showing that externally provided rationales that communicate only a task's extrinsic utility value fail to promote either internalization or engagement in their recipients (Deci et al., 1994; Reeve et al., 2002). Before the extrinsic utility value information contained within a teacher-communicated rationale can be expected to promote either internalization or engagement, it first needs to be embedded within an autonomy-supportive communication style—one that takes the perspective of students, acknowledges their negative feelings, and relies on informational language. To the extent that teachers do not make the instructional effort to help students find personally meaningful connections between their own goals, values, and sense of self and the classroom's uninteresting learning activities, the extrinsic utility value information within a rationale will likely generate only external regulation (not engagement-fostering identified regulation, which is autonomous extrinsic motivation; Reeve et al., 2002). Second, attainment value lies fully within the eyes of the student—in the student's sense of self. Hence, teacher-initiated instructional strategies to promote students' attainment value toward an uninteresting task make little sense. Instead, the most promising engagement-fostering instructional strategy during an uninteresting task appears to be to blend high awareness of what students already value with externally provided rationales to explain the interrelationship between the student's sense of self and the utility value offered by the task. Such an approach characterizes the independent variable used in the present study—namely, using an autonomy-supportive communication style to explain the task's extrinsic utility value to the student's sense of self (as a prospective teacher). This conclusion has potential application not only for promoting value in uninteresting lessons but also for promoting valuing in students' long-term future goals as well (Husman & Lens, 1999; Miller & Brickman, 2004). That is, the effort to promote another's active engagement toward a long-term future goal likely requires both autonomy support to foster self-determined attainment value and explanatory rationales to promote a specific activity's utility value. Third, the findings show a surprising lack of support for the instructional effort to promote interest value during uninteresting activities. Under circumstances in which students are very likely to be extrinsically motivated—because the task itself cannot provide the inherent satisfactions students need to experience intrinsic motivation—IESs were unable to generate the type of motivation students needed to freely engage in and learn from the deeply uninteresting lesson they faced. When tasks are deeply uninteresting, providing rationales in an autonomy-supportive way to promote identified regulation seems to be the more promising engagement-fostering strategy.

### *Future Research*

The findings point to a number of potentially fruitful areas for future research. Rationales produce positive motivational benefits

because they help students perceive the otherwise hidden utility within the uninteresting activity, such as the capacity of the activity to help them develop an important skill. In practice, then, teachers need to communicate to students how they can benefit from the uninteresting tasks that are assigned to them. One question for future research is the extent to which teachers actually attempt to do this. It remains to be studied whether teachers generate the sort of rationales that their students will accept and internalize. Another question to ask is how successful teachers typically are in helping their students internalize the externally provided rationales they hear into self-endorsed reasons of their own. Another question to ask is what qualities of a rationale are most likely to help students accept it as their own. In other words, what qualities allow some rationales to be perceived as convincing and satisfying whereas other rationales are perceived as bogus, empty, or even manipulative? Still another question to ask is whether the effectiveness of a rationale depends somewhat on students' preexisting identified regulation toward the type of activity that includes the target activity. Can a teacher provide an otherwise amotivated student with a convincing and satisfying rationale and see that student begin to experience some level of identified regulation toward that task? Or, does the effectiveness of an externally provided rationale depend on the presence of some preexisting amount of students' identified regulation in that area of study?

Another question for future research is to explore students' capacity to self-generate rationales. It is not at all clear how rare or how commonplace the practice of self-generating a rationale is during academic lessons. It is also an unanswered question whether student-generated rationales might be more motivationally productive than teachers' externally supplied rationales. On the one hand, student-generated rationales would supposedly be richly embedded within their high identified regulation toward the lesson, whereas, on the other hand, a teacher's externally generated rationale would have the advantage and insight of an experienced expert as to what hidden use underlies the enactment of the uninteresting task.

It is also worth clarifying what interest regulation during an uninteresting task is not. IESs to regulate one's interest are not strategies to escape from or avoid uninteresting tasks. An IES does not, for instance, take the student away from the academic task and toward a substitute (and more interesting) alternative, such as daydreaming (off-task IESs). Time spent daydreaming would lower the student's on-task engagement for the academic assignment and hence decrease learning, performance, or skill development. This is why the inclusion of the currently absent behavioral measures of engagement and measures of learning or performance are so important in the conduct of interest regulation research. Future research on interest regulation needs to show that the IES does not undermine performance. Instead, an academically worthwhile IES is one in which the activity to be done remains the same—read the 200-page book, complete the 20 homework questions, or learn the periodic table. What changes is not the lesson but the students' interest while engaged in the lesson. Students have the capacity to create and use IESs, and, in doing so, they experience the benefit of positive emotion and subjective well-being. Future research on interest regulation and IESs might perhaps reveal additional benefits of these strategies. For instance, the large covariation between identified regulation and interest regu-



lation (see Table 2 and Figure 5) raises the possibility that interest regulation, like the provision of a rationale, might function as a facilitating path to enhanced internalization and identified regulation. In pursuing such a research question, future research might investigate the theoretical and conceptual link between IESs and subsequent identified regulation.

A next step in this line of research is to test these hypothesized models using classroom teachers communicating with students, parents communicating with their children, and school counselors communicating with students. In these naturalistic settings, many different rationales are communicated, which raises the question, "What constitutes an effective rationale?" In the present study, I designed the rationale to facilitate two correlated effects: (a) enable the participant to perceive the activity as important enough to become worth one's effort and (b) help the participant make a connection between the activity and a personal goal (i.e., gaining a useful skill in the present study). In this same spirit, Deci and his colleagues (1994) conceptualized an effective rationale as that which is personally meaningful. The phrase "personally meaningful" nicely captures the experience that lies at the intersection of perceived autonomy and perceived importance.

### Limitations

The primary limitation of the present study was that it was not carried out within the context of an on-going classroom environment. Because the rationale was communicated in a written form (to control for extraneous factors), there was no interpersonal relationship between the rationale provider (e.g., a teacher) and its recipients (e.g., students). Within the context of an on-going interpersonal relationship, each of the following qualities in the teacher trying to motivate students has been found to contribute positively to the students' willingness to accept (i.e., internalize) the communicated rationale: warmth (Goodenow, 1993; Midgley, Feldlaufer, & Eccles, 1989), involvement (Skinner & Belmont, 1993), and interpersonal relatedness (Ryan & Powelson, 1991). Presumably, providing a rationale within the context of a warm, caring relationship would produce a stronger effect on students' motivational processes. Thus, the effects obtained in the present study probably underestimate the motivation-enhancing possibilities of teacher-provided explanatory rationales. To test whether this is so, additional research in actual classroom settings is needed. In addition, because only college-age students were used, the applicability of these findings to the younger grade school children remains untested.

### References

- Anderman, E. M., & Maehr, M. L. (1994). Motivation and schooling in the middle grades. *Review of Educational Research*, 64, 287-309.
- Anderson, J. C., & Gerbing, D. W. (1998). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, 103, 411-423.
- Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, 88, 588-606.
- Benware, C., & Deci, E. L. (1984). The quality of learning with an active versus passive motivational set. *American Educational Research Journal*, 21, 775-765.
- Berlyne, D. E. (1966, July 1). Curiosity and exploration. *Science*, 153, 25-33.
- Bollen, K. A., & Long, J. G. (Eds.). (1993). *Testing structural equation models*. Newbury Park, CA: Sage.
- Brophy, J. (1999). Toward a model of the value aspects of motivation in education: Developing appreciation for particular learning domains and activities. *Educational Psychologist*, 34, 75-85.
- Burton, K. D., Lydon, J. E., D'Alessandro, D. U., & Koestner, R. (2006). The differential effects of intrinsic and identified motivation on well-being and performance: Prospective, experimental, and implicit approaches to self-determination theory. *Journal of Personality and Social Psychology*, 91, 750-762.
- Deci, E. L., Eghrari, H., Patrick, B. C., & Leone, D. R. (1994). Facilitating internalization: The self-determination theory perspective. *Journal of Personality*, 62, 119-142.
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.
- Deci, E. L., & Ryan, R. M. (1991). A motivational approach to self: Integration in personality. In R. Dienstbier (Ed.), *Nebraska symposium on motivation: Perspectives on motivation* (Vol. 38, pp. 237-288). Lincoln: University of Nebraska Press.
- DeVolder, M. L., & Lens, W. (1982). Academic achievement and future time perspective as a cognitive-motivational concept. *Journal of Personality and Social Psychology*, 42, 566-571.
- Eccles, J. S., & Midgley, C. (1989). Stage-environment fit: Developmentally appropriate classrooms for young adolescents. In C. Ames & R. Ames (Eds.), *Research on motivation in education: Goals and cognitions* (Vol. 3, pp. 139-186). New York: Academic Press.
- Eccles, J. S., & Midgley, C. (1990). Changes in academic motivation and self-perceptions during early adolescence. In R. Montemayor, G. R. Adams, & T. P. Gullotta (Eds.), *Advances in adolescent development: From childhood to adolescence* (Vol. 2, pp. 134-155). Newbury Park, CA: Sage.
- Eccles, J. S., & Wigfield, A. (1995). In the mind of the achiever: The structure of adolescents' academic achievement related-beliefs and self-perceptions. *Personality and Social Psychology Bulletin*, 21, 215-225.
- Frederick, J. G., & Wallnau, L. B. (2002). *Statistics for the behavioral sciences*. St. Paul, MN: West.
- Goodenow, C. (1993). The psychological sense of school membership among adolescents: Scale development and educational correlates. *Psychology in the Schools*, 30, 79-90.
- Goodlad, J. I. (1984). *A place called school: Prospects for the future*. New York: McGraw-Hill.
- Green-Demers, I., Pelletier, L. G., Stewart, D. G., & Gushue, N. R. (1998). Coping with the less interesting aspects of training: Toward a model of interest and motivation enhancement in individual sports. *Basic and Applied Social Psychology*, 20, 251-261.
- Grolnick, W. S., & Ryan, R. M. (1987). Autonomy in children's learning: An experimental and individual difference investigation. *Journal of Personality and Social Psychology*, 52, 890-898.
- Haladyna, T., & Thomas, G. (1979). The attitudes of elementary school children toward school and subject matters. *Journal of Experimental Education*, 48, 18-23.
- Harter, S. (1981). A new self-report scale of intrinsic versus extrinsic orientation in the classroom: Motivational and informational components. *Developmental Psychology*, 17, 300-312.
- Harter, S. (1982). The perceived competence scale for children. *Child Development*, 53, 87-97.
- Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace.
- Hidi, S., & Harackiewicz, J. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research*, 70, 151-179.
- Hopkins, K., Glass, G., & Hopkins, G. R. (1987). *Basic statistics for the behavioral sciences*. Englewood Cliffs, NJ: Prentice-Hall.

- Hu, L., & Benter, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55.
- Husman, J., & Lens, W. (1999). The role of the future in student motivation. *Educational Psychologist*, 34, 113-125.
- Isaac, J. D., Sansone, C., & Smith, J. L. (1999). Other people as a source of interest in an activity. *Journal of Experimental Social Psychology*, 35, 239-265.
- Jöreskog, K., & Sörbom, D. (2001). *LISREL 8.51: Structural equation modeling with the SIMPLIS command language*. Hillsdale, NJ: Scientific Software International.
- Joussemet, M., Koestner, R., Lekes, N., & Landry, R. (2005). A longitudinal study of the relationship of maternal autonomy support to children's adjustment and achievement in school. *Journal of Personality*, 73, 1215-1235.
- Legault, L., Green-Demers, I., & Pelletier, L. (2006). Why do high school students lack motivation in the classroom? Toward an understanding of academic amotivation and the role of social support. *Journal of Educational Psychology*, 98, 567-582.
- Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis: Effects of sample size. *Psychological Bulletin*, 103, 391-411.
- Meece, J. L., Wigfield, A., & Eccles, J. S. (1990). Predictors of math anxiety and its influence on young adolescents' course enrollment intentions and performance in mathematics. *Journal of Educational psychology*, 82, 60-70.
- Midgley, C., Feldlaufer, H., & Eccles, J. S. (1989). Student/teacher relations and attitudes toward mathematics before and after the transition to junior high school. *Child Development*, 60, 981-992.
- Miller, R. B., & Brickman, S. J. (2004). A model of future-oriented motivation and self-regulation. *Educational Psychology Review*, 16, 9-33.
- Miller, R. B., Greene, B. A., Montalvo, G. P., Ravindran, B., & Nichols, J. D. (1996). Engagement in academic work: The role of learning goals, future consequences, pleasing others and perceived ability. *Contemporary Educational Psychology*, 21, 388-422.
- Minium, E. W., King, B. M., & Bear, G. R. (1993). *Statistical reasoning and psychology in education* (3rd ed.). New York: Wiley.
- Murdock, T. B. (1999). The social context of risk status and motivation predictors of alienation in middle school. *Journal of Educational Psychology*, 91, 62-75.
- Ntoumanis, N., Pensgar, A., Martin, C., & Pipe, K. (2004). An idiographic analysis of amotivation in compulsory school physical education. *Journal of Sport and Exercise Psychology*, 26, 197-214.
- Parker, L. E., & Lepper, M. R. (1992). The effects of fantasy contexts on children's learning and motivation: Making learning more fun. *Journal of Personality and Social Psychology*, 62, 625-633.
- Reeve, J. (1989). The interest-enjoyment distinction in intrinsic motivation. *Motivation and Emotion*, 13, 83-103.
- Reeve, J., Jang, H., Hardre, P., & Omura, M. (2002). Providing a rationale in an autonomy-supportive way as a strategy to motivate others during an uninteresting activity. *Motivation and Emotion*, 26, 183-207.
- Reeve, J., Nix, G., & Hamm, D. (2003). Testing models of the experience of self-determination in intrinsic motivation and the conundrum of choice. *Journal of Educational Psychology*, 95, 375-392.
- Ryan, R. M., & Connell, J. P. (1989). Perceived locus of causality and internalization: Examining reasons for acting in two domains. *Journal of Personality and Social Psychology*, 57, 749-761.
- Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, 55, 68-78.
- Ryan, R. M., & Deci, E. L. (2002). Overview of self-determination theory: An organism dialectical perspective. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 3-33). Rochester, NY: University of Rochester Press.
- Ryan, R. M., & Powelson, C. (1991). Autonomy and relatedness as fundamental to motivation and education. *Journal of Experimental Education*, 60, 49-66.
- Sansone, C., & Smith, J. L. (2000). Self-regulating interest: When, why, and how. In C. Sansone & J. M. Harackiewicz (Eds.), *Intrinsic motivation: Controversies and new directions* (pp. 343-373). New York: Academic.
- Sansone, C., Weir, C., Harpster, L., & Morgan, C. (1992). Once a boring task always a boring task? Interest as a self-regulatory mechanism. *Journal of Personality and Social Psychology*, 63, 379-390.
- Sansone, C., Wiebe, D. J., & Morgan, C. (1999). Self-regulating interest: The moderating role of hardiness and conscientiousness. *Journal of Personality*, 67, 701-733.
- Shell, D. F., & Husman, J. (2001). The multivariate dimensionality of personal control and future time perspective in achievement and studying. *Contemporary Educational Psychology*, 26, 481-506.
- Skinner, E. A., & Belmont, M. J. (1993). Motivation in the classroom: Reciprocal effects of teacher behavior and student engagement across the school year. *Journal of Educational Psychology*, 85, 571-581.
- Vallerand, R. J., Fortier, M. S., & Guay, F. (1997). Self-determination and persistence in a real-life setting: Toward a motivational model of high school dropout. *Journal of Personality and Social Psychology*, 72, 1161-1176.
- Vallerand, R. J., Pelletier, L. G., Blais, M. R., Briere, N. M., Senecal, C., & Vallieres, E. G. (1993). On the assessment of intrinsic, extrinsic, and amotivation in education: Evidence on the on current and construct validity of the Academic Motivation Scale. *Educational and Psychological Measurement*, 53, 150-172.
- Vansteenkiste, M., Simons, J., Lens, W., Soenens, B., & Matos, L. (2005). Examining the motivational impact of intrinsic versus extrinsic goal framing and autonomy-supportive versus internally controlling communication style on early adolescents' academic achievement. *Child Development*, 2, 483-501.
- Werner, C. M., & Makela, E. (1998). Motivations and behaviors that support recycling. *Journal of Environmental Psychology*, 18, 373-386.
- Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review*, 12, 265-310.
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, 25, 68-81.
- Williams, G. C., Wiener, M. W., Markakis, K. M., Reeve, J., & Deci, E. L. (1994). Medical students' motivation for internal medicine. *Journal of General Internal Medicine*, 9, 327-333.
- Wolters, C. (1998). Self-regulated learning and college students' regulation of motivation. *Journal of Educational Psychology*, 90, 224-235.
- Wolters, C. (1999). The relation between high school students' motivational regulation and their use of learning strategies, effort, and classroom performance. *Learning and Individual Differences*, 11, 281-299.

Received April 21, 2007

Revision received May 6, 2008

Accepted May 28, 2008 ■



# College Seniors' Theory of Their Academic Motivation

Shawn Van Etten  
University of Massachusetts at Dartmouth

Michael Pressley  
Michigan State University

Dennis M. McInerney and Arief Darmanegara Liem  
Nanyang Technological University

College seniors participated in an ethnographic interview study about their academic motivations. It was found that grades and graduation are 2 primary distal target goals that motivate their academic efforts during the senior year. A variety of proximal factors were also reported to affect the seniors' motivation. These factors can be divided into students' internal and external factors. Among the internal factors are student characteristics (e.g., social class, expectations) and student beliefs (e.g., belief about control, belief about learning and mastery), whereas the external factors comprise academic-related factors (i.e., course-, examination-, and assignment-related characteristics, reward, and feedback), social factors (i.e., instructors, family members, and peers), general college environment (i.e., physical environment, academic associations, and internship/volunteer opportunities), and extracurricular activities (i.e., fraternities/sororities and sports participation). These results suggest that there is much to learn about academic motivation during the college years. In particular, there is a need for research employing methodologies other than quantitative, survey-based method that can capture the complexities of motivation during college.

**Keywords:** motivation, strategies/tactics, metacognition, higher education, ethnographic interviewing coupled with the method of constant comparison

Many of the recent studies addressing academic motivation have employed an a priori theoretical approach to the identification of students' motivational constructs and how they regulate the students' achievement related functioning and outcomes (see, e.g., Brophy, 2004; Schunk, Pintrich, & Meece, 2008; Stipek, 2002, for recent reviews). Typically, this has meant that researchers have postulated, in advance, the existence of certain motivational constructs, such as academic goal orientations (Covington, 2000; Elliot, 2005; Kaplan & Maehr, 2007; Pintrich, 2000), attributions of success and failures (Weiner, 2004), self-worth (Covington, 2004), self-efficacy (Bandura, 1997; Schunk & Pajares, 2005), future goals (McInerney, 2004; Nurmi, 1991; Phaet, Andriessen, & Lens, 2004; Simons, Vansteenkiste, Lens, & Lacante, 2004), and task value and expectancy for success (Eccles & Wigfield, 1995; Wigfield & Eccles, 2000; Wigfield, Tonks, & Eccles, 2004), and they have attempted to validate these preconceived constructs

through the use of psychometric research techniques, including testing the hypothesized relationships among constructs (Bong, 1996; Pintrich, Conley, & Kempler, 2003). Such an approach to investigating students' academic motivation may, however, (a) constrain researchers from studying other important motivational constructs and their interactions with external, environment factors and (b) fail to elucidate the relationships of these constructs, both among the constructs themselves and between the constructs and their cognitive, behavioral, and affective concomitants (Pintrich, 2000, 2003; Pintrich et al., 2003). In other words, such quantitative investigations may misrepresent the complexity and dynamism of students' academic motivation.

The present study aims to map out the macrovariables potentially affecting academic motivation of college seniors on a day-to-day basis. Since there has been a long history of research investigating academic motivation, we come to this study with certain theoretical sensitivities about the nature of academic motivation in college students emanating from the extant literature that goes back some 40 years ago (e.g., Becker, Geer, & Hughes, 1968, 1995). That literature also makes apparent, however, that there is an important theory-generating role for qualitative research (Strauss & Corbin, 1990, 1997), including in the area of academic motivation of college students, in particular, in identifying motivational variables that are salient in their lives that might have previously been overlooked by researchers.

---

Shawn Van Etten, Department of Institutional Research, University of Massachusetts at Dartmouth; Michael Pressley, Department of Teacher Education, Michigan State University; Dennis M. McInerney and Arief Darmanegara Liem, Centre for Research in Pedagogy and Practice, National Institute of Education, Nanyang Technological University, Singapore.

Michael Pressley died in May 2006. We gratefully acknowledge his contributions to the research.

Correspondence concerning this article should be addressed to Dennis M. McInerney, Centre for Research in Pedagogy and Practice, National Institute of Education, Nanyang Technological University, NIE 02-02-18, 1 Nanyang Walk, Singapore 637616. E-mail: dennis.mcinerney@nie.edu.sg

## Theoretical Sensitivities

As early as the late 1950s, sociologist Howard Becker and his associates conducted a landmark study on the nature of undergrad-

uate studentship (Becker et al., 1968, 1995). An overarching goal for their study was to understand the college experience from the perspective of the students. Becker et al.'s (1968, 1995) primary methodology was participant observation, which meant in that study that the investigators followed students around for long periods of time, observing what they did and asking them questions about the meanings of the actions observed.

One of Becker et al.'s (1968, 1995) major conclusions was that the grade point perspective predominated in the thinking of undergraduate students—that the students were constantly concerned about their grades, for example, fearing that low grades would exclude them from professional school or prevent them from obtaining the best available jobs. The grade point perspective documented by Becker et al. comprised seven subperspectives (Becker et al., 1995, p. 34):

- (1) . . . rewards that students desire cannot be achieved without sufficiently high grades; (2) a successful student . . . will "do well" in his [or her] academic work; (3) doing well in academic work . . . consists of getting a "good" grade point average; (4) . . . where the actions necessitated by the pursuit of grades conflict with other interests, the latter must be sacrificed; (5) to be successful, a student should do whatever is necessary to get "good" grades, not expending effort on any other goal . . . until that has been achieved; (6) . . . grades can be used to judge the personal worth of other students and of oneself; (7) faculty members may be judged . . . according to how difficult they make it to achieve adequate or "good" grades.

In short, Becker et al. (1968, 1995) viewed grades as the "chief institutionalized valuable of the college" (p. 55). It was true that in their study some students thought that learning for its own sake and personal and intellectual development was important, but many more students (indeed most) valued grades above all else. Using a historiographical method, a historian of higher education Horowitz (1987) made the case that grades have been a principal, if not the principal, motivation of American college students at least since the 1700s, and thus supported Becker et al.'s conclusion. Further, the importance of grades as motivators in the eyes of college students was also demonstrated in more recent studies underpinned by different social science perspectives, including those by Abouserie (1994), Michael (1991), Wolters (1998), and Zitzow (1984).

That said, an alternative perspective is that motivation for grades is only one of many variables affecting academic motivation. McKeachie (1961), for example, made the intuitive case that college student motivation depended on many factors, including teaching methods, learning styles and motives, and teaching method by learning style (motive) interactions. It is true that McKeachie agreed that an overarching goal was grades, because they determined in part whether students could attain long-term career goals (i.e., the distal factor of career success was tied to a high grade point average [GPA]). Nonetheless, McKeachie also recognized the power of proximal factors in motivation. He knew that instructors could inspire and motivate, or they could arouse anxiety and uncertainty, undermining student motivation by doing so. McKeachie contended that students analyze the academic tasks they face and make decisions about what and how to study based on the likelihoods of success and the value of the various tasks to them. In short, McKeachie advanced the hypothesis that many other factors than grades play an interactive role in college student academic motivation.

It is important to note, however, we do not imply that Becker et al. (1968, 1995) failed to recognize that there were a variety of factors affecting academic motivation (e.g., extracurricular involvement, the nature of friendships); rather these factors were only covered in passing in their model, with no explanation of just how extracurricular activities and other persons might both encourage and undermine academic motivation. One reason for this was that the participant observation method used by Becker et al. (1968, 1995) involved observation rather than interviewing of students as a central methodology. That is, although questions were posed to participants about the meanings of their behaviors, these seemed not to be as important in driving the conclusions presented in the study as were the observations. Moreover, we infer, based on the salience of the grade point perspective conclusion in the Becker et al. (1968, 1995) write-ups, that this idea emerged as a tentative conclusion early in the study and once identified became a focus of the study (i.e., the investigators were actively attempting to find support for the grade point perspective and were less concerned with elucidating other points that might be made about the undergraduate experience). This suggests that although there was some acknowledgment by Becker et al. (1968, 1995) of multiple factors affecting student motivation, these researchers did not attempt to model completely the motivational world of college students.

In summary, many accounts of academic motivation have focused on specific factors that can influence academic motivation. In general, a number of individual factors have proven to be potent in affecting student academic effort. The real complexity of student motivation, however, cannot come through in investigations that are so limited, for such studies are not intended to be informative about all of the factors affecting or potentially affecting student motivation, but rather they are intended to make the case for the particular influences that are the focus of the investigation. As a result, studies of academic motivation have often considered thin slices of student life rather than attempt to grasp the whole. To study motivation holistically, an inductive, qualitative research approach should be adopted in contrast to the deductive, quantitative approaches typically employed in contemporary studies on motivation.

### *Inductive–Qualitative Research and the Study of College Student Motivation*

In inductive, qualitative research, the researchers do not come to the study with a strong hypothesis but instead intend to induce regularities from participants' perspective. In relation to investigating student academic motivation, the researchers' goal is to identify factors operating in the college environment that might affect academic behavior, rather than to measure factors exactly. The goal is much more to develop a theory, which is very much grounded in data, than to test one (Dey, 2004). The idea was, then, to develop a credible theory of a phenomenon. It is credible because it is based on thorough analyses of data and can be tested in subsequent deductive, quantitative studies. Underpinned by these perspectives and aims, some researchers (e.g., Pressley, Van Etten, Yokoi, Freebern, & Van Meter, 1998; Van Etten, Pressley, Freebern, & Echevarria, 1998) have employed an inductive, qualitative approach in their attempt to capture the full range of variables potentially affecting academic motivation in college stu-



dents. As a first step in that process, these researchers carried out an inductive, qualitative inquiry of academic motivation among college freshmen. The exact method in that study was ethnographic interviewing (Mishler, 1986; Spradley, 1979), with the data analyzed using the method of constant comparison (Strauss & Corbin, 1990, 1997). This recursive process of question generation and response analysis continued until no information gaps were apparent and no novel claims were emerging from the student reports.

More specifically, in Van Etten et al.'s (1998) study, the college freshmen who were the first-round participants responded to the following open-ended questions, with the responses to these questions categorized and related to one another as the beginnings of a grounded theory of freshmen motivation: (a) What is academic motivation? (b) What can enhance or undermine your academic motivation? and (c) Is there anything you would like to add or clarify? After analyzing all participant responses to these questions, the researchers found many information gaps or misconceptions in the emergent theory. As such, new questions (still open-ended but a little more focused) were devised in an attempt to resolve these incongruities. The questions asked in this second round of interviewing included the following: (a) Who or what can enhance or undermine your academic motivation? (b) How can environments affect your academic motivation? (c) What academic activities affect your academic motivation? (d) What non-academic activities affect your academic motivation? (e) How can assignments affect your academic motivation? (f) How can courses or course scheduling affect your academic motivation? (g) How do goals affect your academic motivation? and (h) Are there any additional variables that we have not discussed today, variables that affect your academic motivation? In this study, it took five rounds of questioning by Van Etten and his associates before no new information was entering their model. Table 1 summarizes the major conclusions about freshmen academic motivation that emerged from that study.

*The Present Study*

Building on the past studies reviewed above (Pressley et al., 1998; Van Etten et al., 1998), using an ethnographic interview for data gathering, in the present study we aimed (a) to identify a range of variables/factors, both internal and external to the individual students, that potentially play an influential role in college seniors' academic motivation, (b) to elucidate how these factors may enhance or undermine their motivation, and (c) to build a conceptual model depicting how these potential variables/factors may interact and affect college seniors' motivation in their academic lives. College seniors were the focus in this study for two reasons. First, in relation to the increasing number of dropouts among undergraduate in their senior year of college, the study intended to contribute to the effort of reducing dropouts by providing comprehensive information about the factors that can and do affect motivation at college. Second, complementary to the previous studies of the motivational variables of college freshmen (Pressley et al., 1998; Van Etten et al., 1998), we expected this study to be part of our bid toward the understanding of academic motivation of college students in general. Once factors affecting the freshmen's and seniors' academic motivation are identified, intervention programs developed by the university and stakeholders can then be tailored to the needs of these two distinct groups of students. In our approach to this study, we are aware of and acknowledge the relevance and utility of the many contemporary motivational theories that have been empirically tested, including those major ones mentioned in the beginning of this article (e.g., academic goal orientation theory, attribution theory, expectancy-value theory, future goal perspective, self-worth theory) and that have contributed to researchers' understanding of motivation at all levels. In contrast, and in order to prevent a theoretically biased approach and analysis, we did not begin our investigation with such preconceived motivational constructs theorized in the literature (Dey, 2004). Instead we attempted to be open to all variables/factors potentially affecting academic motivation of the college seniors. In our discussion, however, we will look at and discuss our findings

Table 1  
*Highlights of Freshmen Students' Claims About Academic Motivation*

Motivating factor	Explanation or definition
Good grades	Getting good grades and avoiding bad grades are primary academic motivations.
Other goals in attending college	Getting a good job, independence from family, wanting to meet new people and have fun, and avoiding the real world.
Effort matters	College students are primarily effort theorists, believing effort is key to success.
Personal characteristics of motivated students	They take personal responsibility, feel in control, are organized, monitor progress toward goals, and are not overly anxious. They are good at resisting distractions.
Instructors	Some professors are more motivating than others (e.g., they are interpersonally competent, consistent in their grading, democratic).
Friends and classmates who motivate academic effort	They are enthused about academics and encourage studying.
Family	Families are motivating when they provide encouragement, have realistic expectations, help the student to establish goals, and are confident about academic issues.
Assignments	Challenging but manageable assignments are motivating, as is choice in assignments. Deadlines motivate effort.
Rewards and feedback	Detailed feedback informs and motivates.
General social milieu of the classroom and college	College environments can be generally scholarly, promote responsibility, and provide support, which increases student motivation. In general, smaller classes are more motivating than larger classes.
Physical environment	Environments that motivate studying are well lit, generally quiet, and comfortable.

in light of existing theoretical perspectives of motivation. By doing so, we hope to identify and distinguish between motivational factors/variables that have been explored and documented in the literature and those that have not been widely studied.

### Method

Since the goal of this project was to continue previous studies of undergraduate academic motivation and to map out the full array of college seniors' understandings about academic motivation, we opted for an inductive, qualitative method—the grounded theory approach used in the study of freshmen motivation (see, e.g., Van Etten et al., 1998).

### Participants

The 91 college seniors participating in this study attended an admissions-competitive public university in the northeastern United States. A college senior was defined as having 90 or more college credits and enrolled in a 4-year, degree-granting undergraduate program.

There were four phases of data collection and analyses. The first three phases involved small group interviews that required students to respond to open-ended questions. There were 18 students interviewed during Phase 1, 27 students interviewed during Phase 2, and 29 students interviewed during Phase 3. The Phases 1–3 seniors included 44 women and 30 men ranging from 20 to 36 years of age, with 61 participants ranging from 20 to 23 years of age, and the remaining 13 participants ranging from 24 to 36 years of age. There were 58 Caucasian students, 7 African American students, 5 Latino students, 2 Native American students, and 2 Asian students. Twenty-one students had between 90 and 99 credits, 33 had between 100 and 109 credits, 19 had between 110 and 120 credits, and 1 student had 135 credits. Their college GPAs ranged from 2.14 to 3.92 on a 4-point scale, with 38 students having GPAs between 2.14 and 2.99, and the remaining 36 having GPAs between 3.0 and 3.92. The Phases 1–3 students declared 19 different majors. Thus, the students participating in Phases 1–3 were diverse according to several distinguishing characteristics.

The 17 Phase 4 participants served as a member check for the emergent model (Lincoln & Guba, 1985). There were 14 Caucasian students, 2 African American students, and 1 Asian student. Each had between 90 and 120 college credits, with GPAs ranging from 1.92 to 4.00. These students represented 11 different majors.

### Data Collection, Analyses, and Preliminary Results

Two researchers collected and analyzed all data. There were three interview phases that involved a recursive process of data collection and data analysis, with the results from each phase informing the generation of subsequent interview questions (Lincoln & Guba, 1985; Strauss & Corbin, 1990). Both researchers agreed a priori that phases of data collection would continue until no new information was emerging. Three phases of data collection were necessary to meet this criterion.

*Phases 1–3 data collection and analysis.* All participants were interviewed by two researchers in a quiet room. There were four groups of students interviewed during Phase 1, with 3, 5, 6, and 4 students representing each group. Phase 2 involved interviewing

six groups of students, with 7, 5, 3, 3, 6, and 3 students per respective group. Phase 3 included five groups, with 6, 8, 4, 4, and 7 students per group.

Initially, the two researchers introduced themselves to students, thanked them for participating, and then collected consent and demographic information. All interviews were tape-recorded to allow researchers to clarify notes. At this point, students were informed about the purpose of the research. The initial introduction was as follows,

We are conducting research to identify the variables affecting college seniors' academic motivation. We will ask you several questions. Some of these questions may seem to have common sense answers, but since we do not want to lead your answers, we request that you provide answers with fine details and examples if possible. If you disagree with, or wish to add to another student's response, please feel free to jump into the conversation. We are tape recording this interview so that we can clarify and elaborate on our notes. Please answer openly and honestly. Everything you tell us is completely confidential.

After ensuring that all participants understood the nature of the research, we were ready to ask our first questions. These first questions were very general, intended to elicit a broad array of students' responses that reflect their construal of academic motivation (see Appendix A for the lists of questions asked in Phase 1). Two interviewers were present and took notes summarizing students' responses for each of the four Phase 1 group interviews. Immediately after the first group interview, both interviewers separately analyzed their notes into clauses; these clauses were then synthesized into categories or themes and subcategories. After the first round of coding, to ensure that both interviewers were coding at a comparable level of detail, they exchanged and compared notes. Subsequently, three additional groups were interviewed, with both interviewers agreeing that no new information was emerging at the end of the fourth group interview. The two interviewers again compared categories and subcategories, with clear agreement that at least 14 themes emerged (e.g., goals affecting motivation, feedback affecting motivation, self-characteristics affecting motivation, external factors affecting motivation). Although one of the interviewers coded more clausal statements (i.e., 327 claims) than the second interviewer (i.e., 90 claims), there was clear agreement that comparable results were collected (i.e., the interviewer coding 327 statements recorded very specific claims while the interviewer coding 90 statements tended to make more general claims; thus, most of the 327 claims could be subsumed under the 90 general claims).

At the end of the fourth group interview, there were still many information gaps and emergent information requiring semantic clarification. These concerns were addressed by generating a new set of questions, still open-ended, albeit a bit more focused (see Appendix B for the Phase 2 questions). The only procedural difference between Phase 1 and Phase 2 was that during Phase 2, interviewers followed up less-than-complete student responses with requests for additional information. For example, when a student stated that "Sports really affect my motivation," the researcher asked the student, "How do sports really affect your motivation?"

After each interview, both researchers again coded information into clausal form, with these clauses then related to the preliminary



categories and subcategories emerging from Phase 1 analyses. As new information was collected, tentative categories and subcategories began to merge and/or become reclassified. For example, although feedback as a motivator emerged as a category during Phase 1 coding, it was found to be only a subcategory during Phase 2 coding. That is, during Phase 2, themes such as rewards, grades, and feedback all became subcategories under the main category information about performance.

In contrast to Phase 1 coding, Phase 2 codes had fewer information gaps, and the two interviewers had fewer disagreements requiring resolution. The model was much closer to saturation at this point. Still, there were several issues that required deeper probing. These issues were resolved by generating a third phase of questions, with these questions much more focused than the questions posed in Phases 1 or 2 (see Appendix C for the Phase 3 questions). At the end of the third phase of data collection and analysis, both interviewers agreed that all information gaps and/or misconceptions were resolved. The model was now considered grounded or saturated (see Strauss & Corbin, 1990, 1997), with both interviewers in agreement about the main categories reported by the students and summarized in the Results section of this article.

*Phase 4 data collection and analysis (member checking phase).* Phase 4 involved having 17 students analyze an integrated, comprehensive copy of the claims emerging from Phases 1–3. The purpose of this member check was to ensure that the phenomenology of students' responses was maintained in written form; in addition, this check served to validate the comprehensiveness of the model. The final coded product was 44 double-spaced pages long, with students requested to read and evaluate out loud the accuracy of each claim. In addition, students were requested to write written responses indicating reasons for agreement or disagreement with claims. All student comments served to support the emergent model, including the maintenance of students' interpretations. That is, all of the claims of the emergent model were accepted as correct by these 17 participants.

## Results

### *College Seniors' Model of Academic Motivation*

The seniors reported that there were many variables that can and do affect their academic motivation. The seniors spoke about how the goals of graduating and earning good grades became their primary target goals during their senior year. These goals are distal in nature and inherently future time oriented. Although not entirely mutually exclusive, there were a few large categories of more proximal variables affecting motivation, which can be summarized into two major categories: internal and external factors. Among the internal attributes were students' characteristics (e.g., expectations about themselves, health, past successes and failures, social class) and students' thinking (e.g., beliefs about control, beliefs about whether they are learning much in a course, expectations about courses). The students also cited the influence of a number of factors external to themselves in determining their academic motivation. These included course and assignment characteristics, the nature of examinations and assessments, feedback, rewards, people (i.e., instructors, family members, peers), the college environment (e.g., academic associations, internships and volunteer opportunities, physical characteristics of the environment), and extracurricular participation. Table 2 provides a detailed summary of all of

the individual, conceptually distinct claims made by college seniors about their academic motivation. It is important to state here that the purpose of this research was not to take a tally count of how many times particular themes were mentioned by participants but rather to ensure that the broad sweep of potential motivational influences was given the opportunity to be voiced. This is in line with the agreement by qualitative research methodologists that, depending on the purpose of the study, the outcomes of the ethnographic focus group interview are not necessarily to be quantified (see, e.g., Stewart & Shamdasani, 1998; Wilkinson, 2004). As noted earlier, the primary purpose of this grounded theory study was to generate hypotheses (rather than to test hypotheses) or to address the questions of "what" and "how" or "what is going on" of the college seniors' academic motivation (rather than the issues of "how often" or "how many"). Thus, while we loosely indicate salient motivational influences by comments such as "some students said," this was not based on a precise numerical count. Issues that were occasionally or rarely mentioned were still to be encompassed in our evolving framework. In essence, all claims were given equal weight to generate as rich a data source as possible from which future research may generate hypotheses about student engagement at college. Since visual representations are useful analytical tools in ethnographic research, in particular, to crystallize and chart the consolidated information and to display its line operations (Fetterman, 1998), all of the motivational factors/variables we found and their flow in affecting college seniors' academic motivation are depicted in Figure 1.

### *Academic Goals of Getting Good Grades and Completing the Degree*

As has been true in all previous analyses of college students' academic motivation (e.g., Becker et al., 1968, 1995; Horowitz, 1987), getting good grades was reported as one of the critical determinants or primary target goals of their academic efforts (see Figure 1). The seniors reported that little active processing of academic material would occur without grades. If there were no grades, the seniors reported that they probably would not even attend class. Further, the seniors reported that grades were more motivating if they mattered. For example, there would be greater concern with grades among students expecting to apply to graduate school. The seniors also reported that they were particularly more motivated when graded for personal improvement rather than in competition with other students.

However, a number of situations were noted as undermining the seniors' academic motivation. These included the situations in which the courses they took were too difficult and thus provided little chance to be competitive for the top grades, or in contrast, if students perceived that they were in a course in which all students received high grades. The seniors also stated that their motivation was undermined if they had low-grade expectations and were willing to settle for low or mediocre grades.

Beyond individual grades, however, completing the degree was much on the minds of the seniors, with them offering a variety of reasons or perspectives on how degree completion can affect motivation (see Figure 1). The seniors believed that, in the senior year, most of the hurdles have been surmounted (i.e., most of the

(text continues on page 820)

Table 2

*Summary of College Seniors' Claims About Their Academic Motivation*

College seniors' main goals affecting motivation	
■ <i>How grades and grading policies affect academic motivation:</i>	■ Seniors believe that little active processing of academic material would occur without grades. If there were no grades, they probably would not even attend class.
	<ul style="list-style-type: none"> <li>• Seniors perceive that grades were more motivating if they matter (e.g., to apply to graduate school).</li> <li>• If seniors perceive that they are in a course in which all students receive high grades, there is less academic motivation.</li> <li>• Seniors feel that they are more motivated when graded for personal improvement rather than in competition with other students.</li> <li>• Seniors perceive that in some courses there is little chance to be competitive for the top grades, and this undermines academic motivation.</li> <li>• Some seniors have low grade expectations and are willing to settle for low grades, and this undermines academic motivation.</li> </ul>
■ <i>How completing degree affect academic motivation:</i>	■ Seniors feel that in senior year most of the course-work is completed. Thus, there is no reason to stop when they are so close.
	<ul style="list-style-type: none"> <li>■ All that remains to be completed are electives, allowing students to select courses that are personally interesting or applicable to future occupations.</li> <li>■ Relative to previous coursework, seniors' coursework is often taught from a more applied perspective, not only emphasizing the utility of the material but also hands-on experience (e.g., field placements), and this enhances their motivation.</li> <li>■ Seniors believe that they can complete with minimal time and effort invested, (i.e., they are comfortable at this point in their degree program in missing classes and believe they can superficially process many assignments without much in the way of consequences).</li> <li>■ Seniors believe that if they did not complete, they would have wasted resources (i.e., their parents' money, their time and effort).</li> <li>■ Seniors believe that completing the four-year degree will result in a better job and great earnings in the future, the approval of others, and will be personally satisfying.</li> </ul>
College seniors' internal factors affecting motivation	
Student characteristics	
■ <i>Student characteristics that can positively affect motivation:</i>	<ul style="list-style-type: none"> <li>• Having high expectations and/or goals.</li> <li>• Paying for their own education or having parents pay for education.</li> <li>• Good health.</li> <li>• Feeling successful.</li> <li>• Feeling that they have been recognized for past accomplishment.</li> </ul>
	■ <i>Student characteristics that can negatively affect motivation:</i>
■ <i>Student social class:</i>	<ul style="list-style-type: none"> <li>■ Having goals or expectations that are either very easy or impossible to attain.</li> <li>■ Poor health.</li> <li>■ Wanting to start paying off debts or to have some spending money, resulting in seeking employment or more employment.</li> <li>■ Receiving more recognition for nonacademic pursuits than academic pursuits</li> </ul>
	<ul style="list-style-type: none"> <li>■ Students from lower class backgrounds experience people and institutions (e.g., churches, businesses, drug houses) that can either support or undermine academic motivation.</li> <li>• Students from middle-class backgrounds receive the message that they are "expected to get a college degree" to make a difference in their lives.</li> <li>■ Students from upper-class backgrounds believe that their parents send the message to their children that it is important to get a college degree in order to be perceived as able, that they can make it on their talents.</li> </ul>
Student metacognition	
■ <i>Student beliefs about control:</i>	■ Seniors are more motivated when they believe that have choices in their academic work
	■ <i>Student beliefs about whether they are learning in courses:</i>
■ <i>Expectations about courses and instructors:</i>	■ Motivation is higher when students perceive they are learning something in their courses.
	<ul style="list-style-type: none"> <li>■ Seniors select certain instructors who are more favorable than others in terms of the amount of work they assign, the quality of the work they require, grading policies, instructional approach, and personal qualities, etc.</li> <li>■ Seniors know that certain types of material demand much effort to learn, master, or to get a "good" grade, while other material requires little effort. The expectations about material difficulty can affect whether students take a course.</li> </ul>
■ <i>Student academic planning:</i>	■ Seniors believe that having plans increases academic engagement, recognizing explicitly that academic demands must be addressed during certain times throughout the day or week.
	<ul style="list-style-type: none"> <li>■ A good syllabus in a course is key to student planning. When an informative syllabus is available and stuck with from the beginning of the course, it allows students to gauge the amount of effort and time they will need to complete the course requirements.</li> </ul>
■ <i>Student beliefs about effort, strategy, and ability tradeoffs:</i>	■ While seniors recognize that ability can affect academic performance, they also believe that with effort and efficient strategies, most college students can do well. This belief about "effort-strategy-ability tradeoffs" enhances their motivation.

(table continues)



Table 2 (continued)

## College seniors' external factors affecting motivation

## Academic factors

■ *Course characteristics:*

- Larger class sizes decrease seniors' academic motivation because there is less personal responsibility, classes can be missed without penalty, etc. In contrast, smaller class sizes promote academic motivation because there are more interactions between students and instructors, there is more autonomy, etc.
- Working in large groups often result in students not participating because speaking in front of large groups is difficult for some students. Conversely, smaller groups often are motivating because they feel less threatened.
- When there is group work in a course, it is much more motivating when the students select themselves into groups compared to when instructors assign group members.
- When there is group work in a course, it is much more motivating when all group members complete allocated duties. If some group members fail to complete duties, other group members begin to get worried or disgruntled.
- Study groups can enhance or undermine academic motivation. When all group members are prepared and interested in learning the material, motivation is enhanced. In contrast, when study group members are ill-prepared and not interested in learning the material, motivation is undermined.
- If study group members consist of friends, motivation is often increased, but only if the friends are motivated to learn. However, if study group members do not know one another before studying together, motivation can be enhanced because demands made on members are more often egalitarian.
- Competition in a course can either enhance or undermine motivation. If the competition is against realistic standards (e.g., a "B" student competing against another "B" student, not a "C" student competing against an "A" student), motivation is enhanced. In contrast, when competition sets one up for failure (i.e., it involves competition with much more able students), motivation can be undermined.
- There is greater motivation to do well in courses in the student's major.
- A syllabus that makes the requirements in the course seem overwhelming can undermine motivation.
- Seniors' motivation is low for reading of texts and articles that cannot be understood easily.
- If material in a course is covered in previous courses, there often is little motivation to engage the material again, unless the instructor makes clear the utility and/or applicability of the material in this course.

■ *Assignment characteristics:*

- Seniors are most motivated to do assignments permitting them to apply what they have been learning to a "real" situation.
- Seniors feel that they are motivated to do assignments when they perceive that the material is useful and interesting.
- Choice with respect to assignments increases motivation, although there should not be so many choices that the selection seems overwhelming.
- If the deadline is in the distant future, there is little motivation to do the assignment.
- When large assignments are broken into subtasks distributed across the term (i.e., each subtask has a due date), there is greater motivation than when the large assignment is on one date near the end of the term.
- The difficulty, breadth, and depth of assignments affect student motivation.
- Firm, clear deadlines motivate students to meet those deadlines, whereas the possibility of an extension motivates procrastination.

■ *Examinations and assessments:*

- Seniors recognize that without exams and accountability, much of academic engagement would evaporate. They would read fewer assignments, listen less in class, take fewer notes, and process material less completely.
- Take-home exams can undermine academic motivation when they permit "cheating."
- Some seniors are motivated by take-home exams, believing that take-home exams better capture students' true knowledge than do in-class exams.
- When exams involve group work, there are a variety of potential motivational implications. Motivation can be affected depending on whether there are checks to make certain that all group members contribute to the group product.
- Multiple-choice formats often decrease motivation for seniors because they perceive these formats do not capture student knowledge, which reduces motivation.
- Essay exams motivate seniors in that they allow students to explain themselves. On the other hand, essay exams undermine academic engagement in that students know that often they can write quite a bit even if they know very little, with it easy to skirt around difficult issues in essay responses.
- Graded oral presentations tend to motivate engagement since students believe such presentations are not likely to be received well unless the student is well prepared.
- If a number of exams are occurring at the same time, seniors know to juggle their studying, that is, they know to study less that which is well known and to study more that which is unknown but could be learned given the amount of time available.
- Pop quizzes force students to keep-up with material, but seniors perceive that they cause dislike of the instructor and can undermine motivation or interest toward the course and material.
- It is motivating when professors use exams to diagnose areas of student weakness and then provide opportunities to remediate the weakness.

■ *Feedback:*

- Seniors perceive that positive informational feedback (e.g., this is a good answer because. . .) is the most motivating for them. This type of feedback lets them know that they are doing good work and why.
- Positive evaluative feedback does not have to be informative, however (e.g., good job, correct, good answer), and such feedback is not very helpful.
- If positive informational feedback cannot be given, seniors perceive that the next most motivating type of feedback is negative informational feedback (e.g., this is wrong because. . .). This type of information allows students to correct weaknesses.
- Negative evaluative feedback (e.g. wrong, incorrect, weak) is believed to be least motivating. This type of uninformative feedback often causes students to exert less effort and/or to give-up.

■ *Rewards:*

- Seniors feel that rewards, aside from grades and feedback, can affect their motivation.
- Seniors self-reward themselves when they complete a task, and this enhances their motivation.

Table 2 (continued)

## Social factors

■ *Instructors:*

- Seniors recognize that some instructors are more motivating than others. Some of the following characteristics are shown by the motivating instructors: provide finely-detailed course requirements; relate assignments to course goals; identify critical information for students; present well-prepared lectures; treat all students like adults.

■ *Family members:*

- Family members can enhance seniors' academic coping by providing genuine concern, encouragement, and by helping students establish realistic goals and expectations. In contrast, family members can undermine seniors' academic coping by applying pressure for success, making demands, and by setting unrealistic goals and expectations.

■ *Peers*

- Peers can support academic coping by encouraging studying, by offering an open-minded ear for catharsis, and by being generally academic-oriented. In contrast, peers can undermine academic coping by pressing students to avoid studying and by not providing a source for discussion of academic issues.

## General college environment

■ *Academic associations:*

- Seniors believe that honor societies (e.g., Psi Chi), student councils, and domain-specific associations and clubs (e.g., American Marketing Association, Greek club) stimulate important academic-related learning by providing forums for academic discussions and social networks, that is, they provide students with a way to acquire important academic information.
- One problem with these associations is that for students to be "really accepted" as members, they must devote a lot of personal time (e.g., there is a lot of paper work, scheduling meetings, fund raising events, and other group functions), reducing time for formal academic requirements.

■ *Internships and volunteer opportunities:*

- Seniors think that internship and voluntary activities can enhance students' academic motivation because these activities show why what they are learning is important and is really applicable to their occupation.
- Internships and volunteer opportunities, however, can undermine motivation because these activities require a lot of time, decrease time available for formal academic work, etc.

■ *Physical environment:*

- Seniors perceive that classroom noise, lighting, temperature, cleanliness, and general comfort can affect academic motivation.
- Seniors perceive that motivation can be undermined when a class is in a large classroom as it makes it more difficult for students to see and hear instructors and/or material.
- The weather may affect motivation to attend class, with either beautiful or terrible weather having the potential to reduce motivation to make it to class.
- The distance between home and campus may undermine motivation to attend class.

## Extracurricular activities

■ *Sports:*

- Sports participation can motivate academic activity because most formal college sports require students to maintain certain grade-point-averages and have mandatory study times for lower-achieving students. However, if students become overly devoted to the sport(s), spending a lot of time participating in- or thinking about them, academic motivation can be undermined.

■ *Fraternities and sororities:*

- Fraternities and sororities can motivate academic motivation by requiring participants to maintain certain grade-point-averages, providing a network of individuals knowledgeable about how to do well in college, and having academic competitions within and between fraternities and sororities. Conversely, they can also decrease academic motivation by demanding much time from participants.
- Seniors believe that part-time jobs potentially undermine academic motivation, especially when the job demands much of the students' time.

## College seniors' reasons for working hard, slacking off, and drop out

■ *Reasons for working hard during senior year:*

- Some seniors believe that they always want to do well, and thus, they are continuing to work hard in their remaining coursework.
- Some seniors have experienced a great deal of academic success in college, and this motivates effort during the senior year.
- Seniors, more than other students, are able to select courses that are personally interesting or relevant to future occupational careers, and courses that fit into personal schedules.
- Some seniors perceive that they are really beginning to learn the worth of academic material (e.g., through internships, placements).
- When seniors are still seeking letters of recommendation or believe that remaining courses may affect their acceptance into graduate or professional school, they would exert great effort in the courses that remain.
- By senior year, some students come to the realization that they have had their fun during college (perhaps with some cost in terms of academic achievement), and thus, it is time to really put the effort into academics.
- Some are willing to sacrifice social activities now in favor of academic pursuits, believing that doing well academically now will provide social opportunities for the remainder of their lives.

(table continues)



Table 2 (continued)

- 
- *Reasons for slacking off during senior year:*
    - When seniors already have applied for or been accepted into a graduate or professional training program, academic motivation often decreases because seniors believe that remaining coursework will not affect the decision.
    - If they have already landed a job that will be waiting for them after graduation, there is a similar negative effect on motivation.
    - They believe that the few courses that remain will not much affect overall grade point average.
    - If there are required courses in general education that remain, these seem to be especially unappealing to seniors.
    - Some seniors have experienced a great deal of academic failure in college, and this motivates lack of effort during the senior year. The more failure has occurred following great effort, the less incentive reported to work hard.
    - Seniors invest efforts in keeping the friends they have made. They realize that college will be coming to an end soon and are concerned at solidifying relationships so as to remain in contact with the friends they have made.
    - By senior year, students have developed non-academic interests, (e.g., drawing, shopping, playing with computers, sports).
  - *Motivations to drop out of school (i.e., to not complete the degree):*
    - Some seniors have a history of never achieving academic goals, and thus, they do not expect to complete the degree.
    - Some have fallen so far behind in the specific requirements they need to complete, they perceive there is no hope. Thus, seniors may opt to drop-out rather than spend the overwhelming amount of time and effort that would be required to succeed in advanced courses.
    - Seniors can experience personal emergencies (e.g., injury or serious illness; needing to take care of a sick family member; the death of a close friend or relative).
    - Seniors learn that a college degree will not always secure them a better job (i.e., a better paying job and/or a personally satisfying job).
    - Sometimes seniors learn that a student with 90 or more college credits can often secure the same job as a student with a four-year degree. Seniors reported sometimes being offered jobs requiring them to start immediately.
    - Accepting a job offer before completing a degree permits them to get in better shape financially (e.g., in paying off loans). Students can feel pressed to secure a full-time paying job after having lived off of parents, student loans, and part-time incomes.
    - Seniors want to start living in the "real world." They want to secure the 9-to-5 job that does not require work to be brought home (as compared to college).
    - Seniors want to start applying the information learned over the past 3+ years. They also believe that they will not learn anything new in the last few courses.
    - Seniors perceive that what they lack is experience for the jobs they desire. They perceive that such experience can be gained without having the degree completed.
    - Seniors believe they can complete their degree on a part-time basis while working.
- 

course work is completed). Thus, there was no reason to stop when they were so close. As one senior claimed, "Now it's just a matter of time, or should I say an accumulation of credits?" They felt that if they did not complete, they would have wasted resources (i.e., their parents' money, their time and effort). In relation to the senior year's courses, the seniors espoused that all that remained to be completed were electives, allowing them to select courses that were personally interesting or applicable to future occupations. Further, they stated that, relative to previous coursework, seniors' course work was often taught from a more applied perspective, not only emphasizing the utility of the material but also hands-on experience (e.g., field placements). Thus, seniors believed they could experience a sense of professionalism as they fulfill the requirements that remained. In relation to the consequences of graduating, the seniors perceived that completing the 4-year degree would result in a better job and greater earnings in the future, the approval of others, and personal satisfaction.

### *Student Internal Factors Affecting Academic Motivation*

The seniors in this study reported that students' personal characteristics and their thinking can affect academic motivation. These variables were grouped into one category called *internal Factors* (see Figure 1). In particular, the seniors made observations about the potential role of social class in affecting student motivation. They reported that students from lower class backgrounds experienced people and institutions (e.g., religious institutions, businesses, drug houses) that can either support or undermine academic motivation. Institutions can have a positive influence by supporting the notion that students need a college degree to "get

out of this place," to "help our own," and to "help yourself." They can also have a negative influence on students by suggesting that they have gone far enough in their education. That is, institutional representatives can send the message that a senior from a lower class status has already met and surpassed everyone's expectations, and the senior will always be lower class to everyone else, so why finish the degree.

The seniors in this study also reported that students from middle-class backgrounds received the message that they were expected to get a college degree. The perception was that middle-class students were told that they can make a difference, and hard work resulting in a college degree was what will provide them with opportunities. The seniors also believed that the upper class parents sent the message to their sons and daughters that it is important to get a college degree. The reasons to do so, however, were perceived to be different than for the other two groups. The degree was not so critical to upper class students, who expected a position to be waiting for them in any case, and thus upper class students were reported to be less academically motivated by future employment possibilities. On the other hand, the students reported that some upper class students were motivated to try hard and do well in order to be perceived as able (i.e., that they can make it on their talents).

Among other student characteristics that were reported to affect the seniors' academic motivation were the goals and/or expectations that students set from themselves, with having high goals or expectations positively increasing motivation, whereas having goals or expectations that either are very easy or impossible to attain negatively affecting motivation; the financial resources of

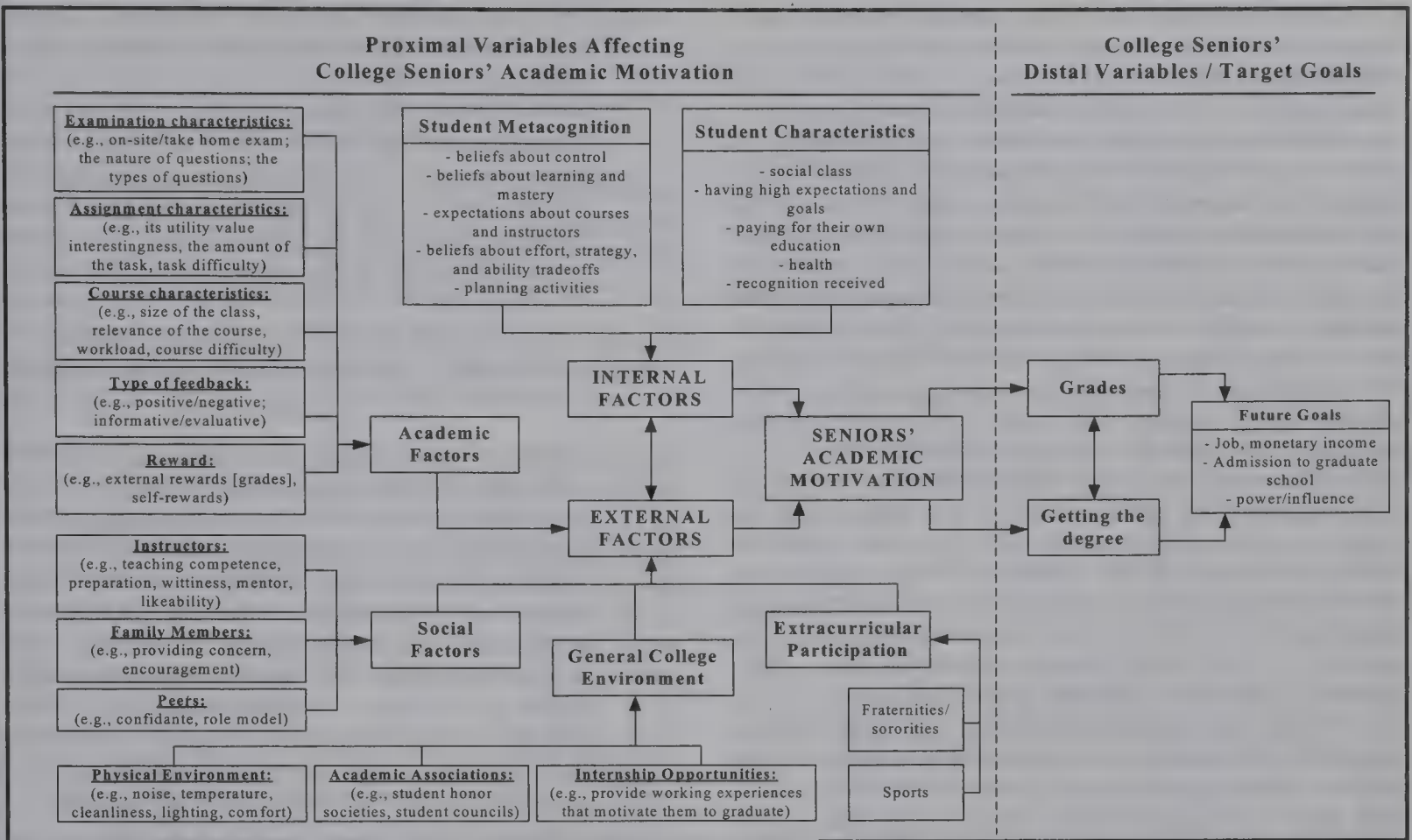


Figure 1. College seniors' model of academic motivation.

their education, with paying for their own education or having parents pay for their education enhancing motivation, whereas wanting to start paying off debts or having some spending money resulting in seeking employment or more employment undermining their academic motivation; health, with good health increasing motivation and poor health decreasing motivation to study; and recognition for past accomplishments, with a feeling or thought that they have been recognized for past accomplishment motivating their academic efforts, whereas receiving more recognition for nonacademic pursuits than academic pursuits potentially demotivating their academic efforts.

The seniors in this study also reported that their thinking was very important in determining academic motivation. This category comprises five different beliefs or thinking process. The first category was student belief about control. The seniors in this study contended that they were more motivated when they believed that they had control or choices in their academic work. The second category was student belief about whether they were learning in courses. Seniors were motivated when they believed that they were learning in their courses. The third category was student expectation about courses and instructors. Based on their experiences in their first three years of college lives, prior to enrolling in courses, seniors often thought about courses and instructors. They have learned that certain instructors are better than others, with instructors evaluated in terms of the amount of work they assign, the quality of the work they require, grading policies, instructional approach, and personal qualities. When possible, seniors attempted to take courses from better instructors, although choice of classes

sometimes was determined by the days and times that classes meet, class size, as well as the number of meeting hours per week. Seniors were also aware that certain types of material demand much effort to learn, to master, or to get a "good" grade in, while other material requires little or less effort. The expectations about material difficulty can affect whether students take a course.

Sometimes students entered classes and confirmed their expectations (i.e., that the instructor was good, the demands were reasonable). At other times, students entered class and expectations were disconfirmed. When expectations were disconfirmed, seniors reported that they often coped with it, for example, by requesting additional time to complete some assignments, handing in work of lower quality, handing assignments in late, or never handing in some assignments. Alternatively, disconfirmation of expectations about teachers and courses can result in a course drop. For example, a student might enroll in a large lecture hall class, expecting minimal personal interaction with the instructor, no attendance policy, and multiple-choice tests. The student has taken this course expecting that it will not be demanding, permitting the student to devote more effort to other courses. If these expectations were not met (e.g., the instructor assigns many group projects that are individually assessed by a teaching assistant), students must adjust their schedules accordingly, the quality of work in their other classes, or drop out of the class. Sometimes dropping out is tempting, rather than risking performance in other courses.

The fourth category was student academic planning. Seniors were active planners before enrolling in classes and throughout the semester. They have learned that in order to succeed in college,



they must plan. Initial planning begins when seniors decided which classes to take, formulating their academic plan in light of other social activities and employment. Seniors created schedules with some flexibility built into them, flexibility allowing them to accommodate for unexpected circumstances (e.g., extra work in a class). They believed that having plans increases academic engagement, recognizing explicitly that academic demands must be addressed during certain times throughout the day or week. Without a fixed routine, students realized that they would be at risk for engaging in anything that comes along when it comes along, often resulting in insufficient time for academics. A good syllabus in a course is key to student planning. When an informative syllabus was available and stuck with from the beginning of the course, it allowed students to gauge the amount of effort and time they would need to complete the course requirements.

The fifth category was student beliefs about effort, strategy, and ability tradeoffs. The seniors recognized that effort exerted and strategies used can affect academic outcomes and that an efficient strategy well matched to task demands can reduce the amount of effort required to carry out an academic task. Although the seniors recognized that ability can affect academic performance, they believed that with effort and efficient strategies, most college students can do well. In particular, they reported that low ability can be offset with efficient strategies and high effort. Still, the seniors were also aware that high effort and/or efficient strategies can still result in repeated failures, with some tasks requiring a great deal of time regardless of efforts and strategies that are used. Even so, in general, the seniors believed that effort and efficient strategies promote success, while little effort and poor strategies typically produce failure.

### *Student External Factors Affecting Academic Motivation*

There are many contextual and social factors affecting student motivation. They are grouped into a first-order category called *external factors* (see Figure 1) and can be divided into four second-order categories: academic factors (course characteristics, assignment characteristics, examination characteristics, feedback, and reward), social factors (instructors, family members, and peers), general college environment (physical environment, academic associations, and internship or volunteer opportunities), and extracurricular participation (fraternities/sororities, sports).

### *Academic Factors*

*Course characteristics.* The seniors reported a number of course-related characteristics that can affect their motivation. These included the following:

1. The size of the class, with larger class size decreasing seniors' academic motivation, and smaller class size promoting their academic motivation. Specifically, seniors perceived that, in large class size, there is less or little personal responsibility. That is, classes can be missed without penalty, and the quality of written assignments can be poor because most often they will be evaluated superficially. Further, in large class size, students can cheat, hand in friends' work, or have friends take tests for them. Large courses usually require only minimal work

(e.g., a midterm and final multiple-choice test). In contrast, the seniors reported that smaller class size promotes academic motivation as there are more interactions between students and instructors, there is usually class discussion requiring students to keep up on the material, there is more autonomy, the evaluation of any assignment is often more detailed in comparison to large class assignments, instructors can account for individual differences and grade for personal improvement, and instructors can assess when a student's work is not consistent with their ability and/or effort. Thus, smaller class size usually requires students to devote more time and effort to academic demands.

2. The size of working groups, with working in large groups potentially undermining academic motivation. Larger groups often result in students not participating because speaking in front of large groups is difficult for some students, they fear being embarrassed in front of peers, and they are never allocated enough time to adequately represent themselves. Conversely, smaller groups often are more motivating, with students reporting that participation increases across the semester. Students reported that this is because they begin to feel more comfortable and less threatened in smaller groups, and they are often allocated sufficient time to fully elaborate responses.
3. The formation of working groups, with student-formed groups more motivating than instructor-formed groups. Further, when there is group work in a course, it is much more motivating when all group members complete allocated duties. If some group members fail to complete duties, other group members begin to get worried or disgruntled.
4. Study groups can enhance or undermine academic motivation. For example, when all group members are prepared and interested in learning the material, motivation is enhanced. In contrast, when study group members are ill prepared or are not interested in learning the material, motivation is often undermined. In addition, if study group members consist of friends, motivation is often increased, but only if the friends are motivated to learn. Even so, if study group members do not know one another before studying together, motivation can be enhanced because demands made on members are more often egalitarian.
5. Type of competition. Competition in a course can either support or undermine motivation. If the competition is against realistic standards (e.g., a B student competing against another B student, not a C student competing against an A student), motivation is enhanced. In contrast, when competition sets one up for failure (i.e., it involves competition with much more able students), motivation can be undermined.
6. There is greater motivation to do well in courses in the student's major.

7. A syllabus that makes the requirements in the course seem overwhelming can undermine motivation.
8. Student motivation is low for reading of texts and articles that cannot be understood easily.
9. If material in a present course was covered in previous courses, there often is little motivation to engage the material again, unless the instructor makes clear the utility and/or applicability of the material in the course.

*Assignment characteristics.* The seniors described a number of assignment-related characteristics that can affect their motivation. These included the following:

1. Applicability value of assignment. Seniors were most motivated to do assignments permitting them to apply what they have been learning to a "real" situation. Seniors believed that they were ready to start applying the definitions, concepts, and theories they were learning. They were ready to start learning about the material in action. For example, they would rather interview people who perform a real-world task than simply read about how to do the task.
2. Utility value of assignment. Seniors reported being motivated to do assignments when they perceived that the material was useful. If assignments or material were perceived to have no utility or no face relevance, motivation was diminished. Seniors reported that they sometimes cheated to complete apparently worthless assignments.
3. Interestingness of assignment. Seniors reported being motivated to do assignments that were perceived as interesting.
4. Difficulty of assignment. The difficulty, breadth, and depth of assignments affected student motivation. Assignments that required too much attention to details or the processing of an overwhelming amount of material were demotivating. If assignments were very difficult (i.e., students have low prior knowledge about the issues, it is unclear what needs to be done to complete the assignment, or the sheer time needed to complete the task would inhibit the completion of other assignments), seniors sometimes sought help. However, regardless of whether they received assistance, they decided either to devote the time necessary to complete the assignment adequately or to devote minimal time and effort to complete the assignment without concern with the quality of the work. What determines how a given student will respond to the difficult assignment? If the student has developed rapport with the instructor or needs something from that instructor (e.g., recommendation), seniors reported being more likely to seek help and attempt to complete the assignment well. Similarly, if the assignment was in their major, or they were very interested in the topic, seniors reported devoting the extra effort. Students reported opting for minimal time and effort when

they did not want to diminish the quality of other assignments they considered more important. However, the seniors also reported that easy assignments did not motivate much effort, unless the student was attempting to impress the instructor or was very interested in the assignment or the material. Easy assignments do not receive much time, unless the time is available. The seniors claimed that the most motivating assignments were challenging but manageable (i.e., with reasonable time and effort, a good quality outcome can be achieved). Instructors can play a key role in conveying to students that assignments are manageable by making clear how the assignment can be accomplished.

5. Availability of choices. Choice with respect to assignments increases motivation, although there should not be so many choices that the selection seems overwhelming.
6. If the deadline is in the distant future, there is little motivation to do the assignment. When large assignments are broken into subtasks distributed across the term (i.e., each subtask has a due date), there is greater motivation than when the large assignment is on one date near the end of the term. The seniors reported that sometimes they broke large assignments into subtasks and set their own deadlines for subtasks in order to increase their motivation for completing the entire task. Further, firm, clear deadlines motivate seniors meeting those deadlines, whereas the possibility of an extension motivates procrastination.

*Examinations characteristics.* The seniors recognized that without exams and accountability, much of academic engagement would evaporate. They would read fewer assignments, listen less in class, take fewer notes, and process material less completely. The following characteristics of examination affect the seniors' academic motivation.

1. Types of questions/tasks in the exam. Seniors reported that multiple-choice formats often decrease motivation for seniors because they perceive that they can usually do well by merely memorizing material. Seniors also perceive that multiple-choice exams often do not capture student knowledge, which reduces motivation. In particular, items that are poorly constructed undermine motivation in that seniors know that some questions can be answered on the basis of structural characteristics of the items rather than requiring knowledge of the course material. In contrast, essay exams motivate students in that they allow students to explain themselves. Nonetheless, essay exams can undermine academic engagement in that students know that often they can write quite a bit even if they know very little, with it easy to skirt around difficult issues in essay responses. Oral presentations that are graded tend to motivate engagement since students believe such presentations are not likely to be received well unless the student is well prepared. Lastly, the seniors reported that far too many exams require only memorization, and this does not motivate academic en-



agement. They perceived exams that require application of what they learn as more motivating but occurring infrequently.

2. Take-home or in-class exams. Seniors believed that take-home exams can undermine academic motivation when they permit cheating. Of course, this negative motivational quality of take-homes can be eliminated (e.g., the exam requires that each student integrates material in a novel way). However, other students are motivated by take-home exams, believing that they better capture students' true knowledge than do in-class exams.
3. Social aspect of exam. When exams involve group work, there are a variety of potential motivational implications. For example, some students like group work more than do other students. Motivation can be affected depending on whether there are checks to make certain that all group members contribute to the group product.
4. The number of simultaneous exams. If a number of exams are occurring at the same time, seniors know to juggle their studying, that is, they know to study less that which is well known and to study more that which is unknown but could be learned given the amount of time available.
5. Aim of exam. Seniors reported that it is motivating when professors use exams to diagnose areas of student weakness and then provide opportunities to remediate the weakness. One way to do so is to permit students to correct incorrect exam answers for some credits.
6. Quizzes. Pop quizzes force students to keep up with material, but the seniors report that they cause dislike of the instructor and can undermine motivation or interest toward the course and material.

*Feedback.* In addition to grades, seniors receive other types of feedback (e.g., verbal feedback about the student's potential for success in the field). With regard to feedback, seniors reported that positive informational feedback (e.g., "This is a good answer because . . .") is the most motivating for them. This type of feedback lets them know that they are doing good work and why. However, positive feedback can simply be evaluative and not informative (e.g., "good job," "correct," "good answer"), and such feedback is not very helpful. If positive informational feedback cannot be given, the seniors reported that the next most motivating type of feedback is negative informational feedback (e.g., "This is wrong because . . ."). This type of information allows students to correct weaknesses. It is not surprising that negative evaluative feedback (e.g., "wrong," "incorrect," "weak") was reported as least motivating. This type of uninformative feedback often causes students to exert less effort and/or to give up.

*Rewards.* Seniors reported that rewards, aside from grades and feedback, can affect their motivation. The students reported that external rewards are rare (e.g., admissions to graduate school, for example, come only to a few and then only at the end of the degree program). In part, because of the lack of external rewards, students self-reward themselves. The seniors reported that for self-rewards

to be effective, self-reward should not occur until the task is complete, self-rewards should be realistic and cost efficient (i.e., students cannot afford to go to the movies every day or buy a new set of clothes each day; thus, self-rewards need to be things such as free time for reading, talking on the phone, or sleeping).

### *Social Factors*

In general, people who are motivating are those who support students' academic engagement, provide strategies for dealing with academic demands, and are willing to listen and help resolve students' personal and academic problems.

*Instructors.* Instructors vary in ways that can affect motivation, with more motivating instructors having the following characteristics and/or practices: They provide finely detailed course requirements; they relate assignments to course goals; they identify critical information for students; they provide old tests and other material to help students review for exams; they test only material covered in the course; they provide adequate time for students to complete assignments and tests; they present well-prepared lectures; they advocate democratic discussion during class; they connect course material to pragmatic issues; they have good elocution skills and speak without an accent; they seem enthusiastic and interested in the material; they are responsible, knowledgeable, and permit choices; they are caring and empathetic to students' personal and academic concerns; they establish realistic but challenging goals; they provide realistic encouragement; they provide timely informational feedback; they grade for personal improvement and emphasize the learning process over grades; they grade based on multiple assessments rather than one big assessment (e.g., instead of one large paper grade, there is grading of subtasks accomplished as the paper is composed); and they treat all students like adults.

*Family members.* The seniors in this study reported that family members can enhance students' academic motivation by providing genuine concern, by providing encouragement, and by helping students establish realistic goals and expectations. In contrast, family members can undermine students' motivation by applying pressure for success, by making demands, and by setting unrealistic goals and expectations.

*Peers.* The seniors in this study also stated that peers can support academic motivation by encouraging studying, by offering an open-minded ear for catharsis, and by being generally academic oriented. In contrast, peers can undermine academic motivation by pressuring students to avoid studying and by not providing a source for discussion of academic issues. The awareness of the role of peers in motivation prompts action by some students, for example, to seek new friends in an attempt to do better academically. Other students, however, reported that they did not seek to control their motivation by affecting the network of people surrounding them.

### *The General College Environment*

Seniors contended that the general college environment can range from promoting responsibility, encouraging effort, and providing support to encouraging dependency, ability, and a lack of support. For example, if faculty, college personnel, and other students in a college or department really attempt to help others



(e.g., by steering them through red tape), this can support academic motivation. Academic environments that are not so supportive require students to do everything for themselves, increasing the likelihood that students will be overwhelmed.

*Academic associations.* The seniors reported that honor societies (e.g., Psi Chi), student councils, and domain-specific associations and clubs (e.g., American Marketing Association, Greek club) stimulate important academic-related learning by providing forums for academic discussions and social networks, that is, they provide students with a way to acquire important academic information. These associations help students build up résumés and locate employment. Sometimes, they also provide help with general (e.g., problems with an instructor) and specific (e.g., seeking information about course material) academic issues. One problem with these associations is that for students to be really accepted as members, they must devote a lot of personal time (e.g., there is a lot of paper work, scheduling meetings, fundraising events, and other group functions), reducing time for formal academic requirements.

*Internships and volunteer opportunities.* Seniors often have internships and/or volunteer. The seniors reported that these activities can enhance students' academic motivation in different ways: They make more obvious why what they are learning is important and is really applicable to their occupation; they provide feedback about whether they can perform the jobs to which they aspire; and they sometimes expose students to opportunities that require post-baccalaureate education, which motivates students to go to graduate or professional school. In contrast, internships and volunteer opportunities can undermine motivation in the following ways: They can demand a lot of time, decreasing time available for formal academic work; they may be too difficult or too easy; the students do not enjoy the applied nature of their field; or the students can learn that they do not have the qualities necessary to fulfill job requirements. In other words, the seniors believed that they learned most and were most motivated by an internship or volunteer opportunity of medium difficulty, one that allows them to learn whether they would prefer such a position in the future but one that does not require so much time that course work would be affected.

*Physical environment.* The students reported that the immediate physical environment has a big impact on whether a student is motivated. The seniors reported that classroom noise, lighting, temperature, cleanliness, and general comfort can affect academic motivation. They reported that motivation can be undermined when a class is in a larger classrooms, which makes it more difficult for students to see and hear instructors and/or material. Further, if the physical environment is filled with people the student knows, it can affect motivation to attend or skip class. For example, if the student is in an environment in which he or she is likely to see friends before class, those friends sometimes provide temptations to skip class. The seniors also reported that weather can affect motivation to attend class, with either beautiful or terrible weather having the potential to reduce motivation to make it to class. And lastly, if a student lives a distance from campus, this can undermine motivation to attend class.

### *Extracurricular Activities*

The seniors in this study also reported that extracurricular activities or participation also have the potential either to enhance or

detract from academic motivation. One of these extracurricular activities is sports participation, which was reported to have a positive effect on the seniors' academic motivation. Most formal college sports require students to maintain certain GPAs and have mandatory study times for lower achieving students. Participation in sports also forces many seniors to structure an otherwise unstructured lifestyle, including scheduling of study. Both formal and informal sports also provide seniors with a break from academics. Seniors perceived that these breaks allow students later to tackle academic demands with renewed vigor. Still, if students become overly devoted to the sport(s), spending a lot of time participating in, or thinking about them, academic motivation can be undermined. Also, sporting events often are scheduled during school hours, which undermines academic motivation by producing the perception that it is acceptable to miss classes and that sports are more important than academics.

Another extracurricular activity reported by the seniors as affecting their academic motivation is fraternities/sororities. The participation in fraternities/sororities can enhance motivation because it requires participants to maintain certain GPAs, providing a network of individuals knowledgeable about how to do well in college and having academic competitions within and between fraternities and sororities. Conversely, they can also decrease academic motivation by demanding much time from participants and providing old tests, papers, and/or other assignments.

### Discussion

Van Etten et al. (1998) made the case that the factors affecting the academic motivation of freshmen were many and complex. In this study, it was clear that the motivational influences on seniors are greater and more complex. For example, college seniors are oriented to two worlds, both the supports and demands of college and the uncertain challenges of the postcollege world (cf., Chickering & Schlossberg, 1998). With respect to the demands and supports of college itself, there was a distinct difference between the seniors and the freshmen we studied previously. Seniors are critically aware that their negotiation of college per se is an end-of-a-venture experience, for example, saliently aware of the competing motivations for completing the degree versus dropping out and for working hard in senior year versus slacking off. The freshmen, of course, were motivated to figure out the challenging, new environment they were confronting.

That seniors have some distinct motivations should not obscure, however, that there also is much overlap in the motivations of the freshmen in Van Etten et al. (1998) and the seniors in this study. For example, both groups of students cited a variety of personal characteristics as important in determining motivation, although again there were many more characteristics cited by the seniors, including a keen awareness that the nature and reasons for academic motivation can vary with social class. Freshmen and seniors also know that student thinking really matters, although again, experience in college permitted much more detailed reports by seniors about how thinking affects motivation than by freshmen (e.g., seniors have detailed expectations about courses based on what they know about instructors and other students' experiences, and they know how they react to violations of their expectations). A related point is that although freshmen knew in general that course, assignment, and instructor characteristics matter, the se-



niors offered detailed observations about how courses, assignments, exams, and instructors can vary in ways that affect motivation. Although freshmen seem to know that the general college environment affected their motivation and could provide detailed information about some aspects of the environment and its effects on motivation (e.g., qualities of the physical environment), the general environment of the seniors is richer and more complex. For example, the seniors have experienced the effects of departments that are helpful to students versus those that are not and know how the department atmosphere can affect motivation. Seniors have also participated in the interactions that occur in clubs, internships, sports, and jobs and know how these aspects of college life can positively and negatively affect academic motivation.

One very salient similarity between freshmen and seniors, and one confirming the perspective development by Becker et al. (1968, 1995), is that grades are important in the thinking of college students. That said, the seniors provided much more information about occasions in college when getting a good grade matters less and when it matters more. Although a good case could be made from the Van Etten et al. (1998) data that getting good grades is the main orientation of freshmen, with seniors, the conclusion has to be adjusted: Getting good grades is often important to seniors, although seniors also know when grades just do not matter that much (e.g., when their graduate school applications and accompanying transcripts are in the mail at midsenior year).

Something that was very intriguing in the senior results, when considered in light of the motivational literature in general, is that the senior students reported in several different ways that neither easy nor overly difficult tasks were motivating. Rather, tasks that are moderately difficult are the most engaging, consistent with theoretical perspectives as diverse as White's (1959) competence model and Vygotsky's (1978) conception of the zone of proximal development (i.e., when tasks are just a bit beyond what the student can do already, students are likely to be most motivated).

One great value of a study such as this one is that it provides so much information to those who serve college students—such as professors and administrators—about what they can do to motivate students and, conversely, the things that can happen in college that undermine student motivation. Indeed, we are struck that this interview study provided much more expansive information than the Van Meter, Yokoi, and Pressley (1994) study about factors that make a difference in facilitating versus undermining students' learning and academic motivation. In that investigation, the focus was on student note taking and the factors affecting it, which was expected to be illuminating since note taking is the principal activity occurring in college class meetings. What becomes clear in this study as well as in Van Etten et al.'s (1998) study is that the factors affecting note taking are just a small subset of the many factors defining student motivation. Motivation is determined by in-class factors and out-of-class factors, by instructor characteristics, but also by the characteristics of many others in the college environment. Consistent with Van Meter et al.'s conclusions, however, what emerges from comparing the freshmen results in Van Etten et al.'s (1998) study and these senior results is that freshmen and seniors are very different, with seniors having highly differentiated and complex conceptions of the college world compared to freshmen. Colleges are complex places, and 4 years of studentship reveal many complexities that are important to students.

An obvious criticism of this study is that it relied on interview data, rather than ethnographic observations. Although there have been few ethnographic observational studies of the college experience, and none that we know of that focus on academic motivation per se, there is reason to believe that when an ethnographic observational study is carried out that focuses on motivation, the results will complement the interview outcomes presented here. For example, in reading anthropologist Michael Moffatt's (1989) *Coming of Age in New Jersey: College and American Culture*, an ethnographic observational study of college life conducted in the late 1970s and early 1980s at Rutgers, we found many bits and pieces of information confirming the big motivational categories detected in this study and in Van Etten et al.'s (1998) study. Notably, however, the extensive detail reported in this study was missing with respect to motivation in Moffatt's report, of course, reflecting that academic motivation was not the central concern in that study. Even so, there is every reason to believe based on Moffatt's study that observation of students will yield information about the importance of grades and long-term goals; information about how and when students exert academic efforts; insights about how professors, dorm mates, and parents affect student course taking and studying; and how the many aspects of the college environment affect academic motivation. Even more optimistically, perhaps with this work providing theoretical sensitivity to the future ethnographic observer, that work will be even more informative than it would be otherwise. The richness of the academic motivations of seniors suggested by the results reported here, and the richness of the academic motivations of freshmen as suggested by Van Etten et al. (1998), provide plenty of motivation for researchers to continue to study the motivations of college students in a more structured manner.

## References

- Abouserie, R. (1994). Sources and levels of stress in relation to locus of control and self esteem in university students. *Educational Psychology, 14*, 323–330.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.
- Becker, H. S., Geer, B., & Hughes, E. C. (1968). *Making the grade: The academic side of college life*. New York: Wiley.
- Becker, H. S., Geer, B., & Hughes, E. C. (1995). *Making the grade: The academic side of college life*. New Brunswick, NJ: Transaction.
- Bong, M. (1996). Problems in academic motivation research and advantages and disadvantages of their solutions. *Contemporary Educational Psychology, 21*, 149–165.
- Brophy, J. (2004). *Motivating students to learn*. Mahwah, NJ: Erlbaum.
- Chickering, A. W., & Schlossberg, N. K. (1998). Moving on: Seniors as people in transition. In J. N. Gardner, G. Van der Meer, & Associates (Eds.), *The senior year experience: Facilitating integration, reflection, closure, and transition* (pp. 37–50). San Francisco: Jossey-Bass.
- Covington, M. V. (2000). Goal theory, motivation, and school achievement: An integrative review. *Annual Review of Psychology, 51*, 171–200.
- Covington, M. V. (2004). Self-worth theory: Goes to college or do our motivation theories motivate? In D. M. McInerney & S. Van Etten (Eds.), *Big theories revisited* (pp. 91–114). Greenwich, CT: Information Age.
- Dey, I. (2004). Grounded theory. In C. Seale, G. Gobo, J. F. Gubrium, & D. Silverman (Eds.), *Qualitative research* (pp. 80–93). London: Sage.
- Eccles, J., & Wigfield, A. (1995). In the mind of the actor: The structure

- of adolescents' achievement task values and expectancy-related beliefs. *Personality and Social Psychology Bulletin*, 21, 215–225.
- Elliot, A. J. (2005). A conceptual history of the achievement goal structure. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 52–72). New York: Guilford Press.
- Fetterman, D. M. (1998). Ethnography. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 473–504). London: Sage.
- Horowitz, H. L. (1987). *Campus life: Undergraduate cultures from the end of the eighteenth century to the present*. New York: Knopf.
- Kaplan, A., & Maehr, M. L. (2007). The contributions and prospects of goal orientation theory. *Educational Psychology Review*, 19, 141–184.
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.
- McInerney, D. M. (2004). A discussion of future time perspective. *Educational Psychology Review*, 16, 141–151.
- McKeachie, W. J. (1961). Motivation, teaching methods, and college learning. In M. R. Jones (Ed.), *Nebraska symposium on motivation* (pp. 111–142). Lincoln: University of Nebraska Press.
- Michael, J. (1991). A behavioral perspective on college teaching. *Behavior Analyst*, 14, 229–239.
- Mishler, E. G. (1986). *Research interviewing: Context and narrative*. Cambridge, MA: Harvard University Press.
- Moffatt, M. (1989). *Coming of age in New Jersey: College and American culture*. New Brunswick, NJ: Rutgers University Press.
- Nurmi, J.-E. (1991). How do adolescents see their future? A review of the development of future orientation and planning. *Developmental Review*, 11, 1–59.
- Phalet, K., Andriessen, I., & Lens, W. (2004). How future goals enhance motivation and learning in multicultural classrooms. *Educational Psychology Review*, 16, 59–89.
- Pintrich, P. R. (2000). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology*, 92, 544–555.
- Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of Educational Psychology*, 95, 667–686.
- Pintrich, P. R., Conley, A. M., & Kempler, T. M. (2003). Current issues in achievement goal theory and research. *International Journal of Educational Research*, 39, 319–337.
- Pressley, M., Van Etten, S., Yokoi, L., Freebern, G., & Van Meter, P. (1998). In D. J. Hasher, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 347–366). Mahwah, NJ: Erlbaum.
- Schunk, D. H., & Pajares, F. (2005). Competence perceptions and academic functioning. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 85–104). New York: Guilford Press.
- Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2008). *Motivation in education: Theory, research, and applications* (3rd ed.). Upper Saddle River, NJ: Pearson Prentice Hall.
- Simons, J., Vansteenkiste, M., Lens, W., & Lacante, M. (2004). Placing motivation and future time perspective theory in a temporal perspective. *Educational Psychology Review*, 16, 121–139.
- Spradley, J. P. (1979). *The ethnographic interview*. New York: Holt, Rinehart, & Winston.
- Stewart, D. W., & Shamdasani, P. N. (1998). Focus group research: Exploration and discovery. In L. Bickman & D. J. Rog (Eds.), *Handbook of applied social research methods* (pp. 505–526). London: Sage.
- Stipek, D. (2002). *Motivation to learn: Integrating theory and practice* (4th ed.). Boston: Allyn & Bacon.
- Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage.
- Strauss, A., Corbin, J. (1997). *Grounded theory in practice*. London: Sage.
- Van Etten, S., Pressley, M., Freebern, G., & Echevarria, M. (1998). An interview study of college freshmen's academic motivation. *European Journal of Psychology and Education*, 13, 105–130.
- Van Meter, P., Yokoi, L., & Pressley, M. (1994). College students' theory of note-taking derived from their perceptions of note-taking. *Journal of Educational Psychology*, 86, 323–338.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Weiner, B. (2004). Attribution theory revisited: Transforming cultural plurality into theoretical unity. In D. M. McInerney & S. Van Etten (Eds.), *Big theories revisited* (pp. 13–29). Greenwich, CT: Information Age.
- White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review*, 66, 297–333.
- Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology*, 25, 68–81.
- Wigfield, A., Tonks, S., & Eccles, J. S. (2004). Expectancy-value theory in cross-cultural perspective. In D. M. McInerney & S. V. Etten (Eds.), *Big theories revisited* (pp. 165–198). Greenwich, CT: Information Age.
- Wilkinson, S. (2004). Focus group research. In D. Silverman (Ed.), *Qualitative research: Theory, method and practice* (2nd ed.) (pp. 177–199). London: Sage.
- Wolters, C. A. (1998). Self-regulated learning and college students' regulation of motivation. *Journal of Educational Psychology*, 90, 224–235.
- Zitzow, D. (1984). The college adjustment rating scale. *Journal of College Student Personnel*, 25, 160–164.

(Appendixes follow)



## Appendix A

## The List of the Phase 1 Questions

1. What is academic motivation?
2. When is your academic motivation affected?
3. Where is your academic motivation affected?
4. Why is your academic motivation affected?

5. How is your academic motivation affected? *Note.* The presentation order of Questions 2–5 was varied for each group, as we felt this would stimulate the maximum variability in responses.

## Appendix B

## The List of the Phase 2 Questions That Were Generated to Clarify Phase 1 Misconceptions and Information Gaps

1. What is enhancing or undermining your motivation to continue in college, and how?
2. What enhances or undermines your motivation to do well or succeed in college, and how?
3. Who enhances or undermines your motivation to do well in college, and how?
4. How can the environment where you learn enhance or undermine your motivation to do well academically?
5. What academic activities can enhance or undermine your motivation to do well academically?

6. What nonacademic activities can enhance or undermine your motivation to do well academically?
7. How can assignments or class material enhance or undermine your motivation to do well academically?
8. Are there any additional variables affecting your academic motivation that we have not discussed today? Are there any other things you would like to add or clarify?

## Appendix C

## The List of the Phase 3 Questions

1. What is boring and/or interesting about being a college senior? Does this affect your academic motivation? How?
2. What is different about being a college senior versus your first three years of college? Does this affect your academic motivation? How?
3. Are there certain activities that college seniors engage in that can affect academic motivation?
4. As seniors, is the material you discuss or how you discuss it affect your academic motivation? Is this different from your first three years?
5. As seniors, what results in superficial or more elaborate processing of academic material?

6. Do some college seniors cheat? What are some of the reasons for cheating as a college senior? How do they cheat?
7. Are there any similarities and/or differences about the “real world” and life as a college senior? How do these similarities or differences affect your academic motivation?
8. What does it take to become a college senior?
9. Do you have anything else you would like to add or clarify, questions that you think we should be asking?

Received April 23, 2007

Revision received February 5, 2008

Accepted February 7, 2008 ■

# Dynamic Assessment of Algebraic Learning in Predicting Third Graders' Development of Mathematical Problem Solving

Lynn S. Fuchs, Donald L. Compton, Douglas Fuchs, Kurstin N. Hollenbeck, Caitlin F. Craddock, and  
Carol L. Hamlett  
Vanderbilt University

Dynamic assessment (DA) involves helping students learn a task and indexing responsiveness to that instruction as a measure of learning potential. The purpose of this study was to explore the utility of a DA of algebraic learning in predicting third graders' development of mathematics problem solving. In the fall, 122 third-grade students were assessed on language, nonverbal reasoning, attentive behavior, calculations, word-problem skill, and DA. On the basis of random assignment, students received 16 weeks of validated instruction on word problems or received 16 weeks of conventional instruction on word problems. Then, students were assessed on word-problem measures proximal and distal to instruction. Structural equation measurement models showed that DA measured a distinct dimension of pretreatment ability and that proximal and distal word-problem measures were needed to account for outcome. Structural equation modeling showed that instruction (conventional vs. validated) and pretreatment calculation skills were sufficient to account for math word-problem outcome proximal to instruction; by contrast, language, pretreatment word-problem skill, and DA were needed to forecast learning on word-problem outcomes more distal to instruction. Findings are discussed in terms of responsiveness-to-intervention models for preventing and identifying learning disabilities.

**Keywords:** dynamic assessment, math word problems, responsiveness to intervention

A major purpose of educational assessment is to forecast academic achievement. The goal is early identification of students who are at risk for poor learning outcomes so that intervention can be initiated before the development of severe academic deficits, which can be intractable and can create life-long difficulty in and out of school (e.g., Rivera-Batiz, 1992). The predominant approach for forecasting academic achievement is traditional testing with a general measure of intelligence (e.g., Raven Progressive Matrices; Raven, 1960) or a test of specific ability or skill presumed to underlie future academic performance (e.g., phonological processing for development of word-reading performance or calculations skill for development of mathematics word-problem performance). In these conventional testing situations, examinees respond without examiner assistance, and a body of work demonstrates that assessments of intelligence or precursor abilities/skills capture varying amounts of variance in forecasting academic development. For example, a measure of quantity discrimination, when used as a screener near the beginning of first grade, accounts for 25%–63% of the variance in end-of-year math outcomes

depending on the study (e.g., Chard et al. 2005; Clarke & Shinn, 2004; Lembke & Foegen, 2006).

Because these conventional assessments are imperfect predictors of academic learning (e.g., Sternberg, 1996), they have long been the target of scrutiny and criticism (e.g., Tzuriel & Haywood, 1992). A major concern is that these “static” estimates of performance reveal only two states: unaided success or failure. By contrast, as Vygotsky (1934/1962) proposed, children may function somewhere between these states: unable to perform a task independently but able to succeed with assistance. This has implications for discriminating students at the lower end of the distribution. For example, when two children earn the same low score on a calculations test, they may not have the same potential to develop word-problem skill. One may succeed in learning how to solve word problems given only minimal assistance. This would suggest that the initially low performance on the static assessment stems from inadequate learning opportunity in the child's present environment but indicating good learning potential with competent instruction in the future. The other child may struggle to learn to solve word problems even when provided highly explicit instruction (revealing the need for special intervention).

So the question arises: If the goal is to forecast learning potential, why not assess the student's capacity to learn, rather than assessing what the student presently knows? This alternative form of assessment, whereby students' learning potential is measured, is known as *dynamic assessment* (DA). DA has been the focus of discussions and research for more than 75 years (e.g., Kern, 1930; Penrose, 1934; Rey, 1934, as cited in Grigorenko & Sternberg, 1998). In the present study, we considered the contribution of a DA in forecasting students' development of word-problem com-

---

Lynn S. Fuchs, Donald L. Compton, Douglas Fuchs, Kurstin N. Hollenbeck, Caitlin F. Craddock, and Carol L. Hamlett, Department of Special Education, Vanderbilt University.

This research was supported in part by National Institute of Child Health and Human Development Grant RO1 HD46154 and Core Grant HD15052 to Vanderbilt University. Statements do not reflect agency position or policy, and no official endorsement should be inferred.

Correspondence concerning this article should be addressed to Lynn S. Fuchs, 228 Peabody, Vanderbilt University, Nashville, TN 37203. E-mail: lynn.fuchs@vanderbilt.edu



petence across third grade. In this introduction, we present a framework for considering prior DA work related to the present study. Then we clarify how the present study builds on and extends this literature.

### Prior Work on DA as a Predictor of Academic Development

DA involves structuring a learning task, providing feedback or instruction to help the student learn the task, and indexing responsiveness to the assisted learning phase as a measure of learning potential. Research on DA varies as a function of the structure and design of the DA and in terms of the methodological features of the studies.

In terms of structure and design, DAs vary along three dimensions (see Campione, 1989): *index*, style of interaction, and the nature of the skills assessed. *Index* refers to the way in which DAs quantify responsiveness to the assisted phase of learning. This is the measure of learning potential. The first strategy for indexing performance is to characterize the amount of change from an unassisted pretest to an unassisted posttest (with the assisted learning phase intervening between the pre- and posttest) or by scoring students' unaided performance following the assisted phase of assessment (e.g., Ferrara, Brown, & Campione, 1986). The second approach for indexing performance is to quantify the amount of scaffolding required during the assisted phase of assessment to reach criterion performance (e.g., Murray, Smith, & Murray, 2000; Spector, 1992). These alternative methods for indexing DA performance serve the same purpose: to predict whether students require extra attention in order to learn adequately. Researchers who use change from unassisted pre- to posttest or use unaided performance following the assisted phase of assessment typically contrast normal learners against special populations of learners (e.g., students with mental retardation) to examine whether the assisted learning experience produces differential learning outcomes as a function of having a diagnosis associated with inadequate learning (e.g., mental retardation). This suggests that low DA scores (less improvement as a function of assisted learning or unaided performance score following assisted learning) serve to predict poor outcomes in other learning situations, indicating that extra instructional support is required to produce adequate learning. In a related way, when research shows that amount of scaffolding predicts learning outcomes outside of DA, this provides evidence that low DA scores (a lot of scaffolding needed to reach criterion performance during the assisted learning phase) serve to predict poor outcomes in other learning situations, indicating that extra instructional support is required to produce adequate learning.

The second dimension along which DA varies is the style of interaction. Some DAs (e.g., Ferrara et al., 1986) are standardized, where the tester administers a fixed series of prompts; success with early prompts reflects the need for minimal adult intervention to perform/learn the task, whereas success on later prompts reflects the need for more extensive adult help to perform/learn the task. By contrast, other DAs (e.g., Tzuriel & Feuerstein, 1992) are individualized, where the tester addresses the student's specific obstacles as revealed by that student's responses.

The third dimension along which DA varies is the nature of the skills assessed. Early work (e.g., Budoff, 1967; Feuerstein, 1979) tended to focus on domain-general skills associated with cognitive ability. In more recent work, tasks tend to be more academically grounded (e.g., Bransford, Delclos, Vye, Burns, & Hasselbring, 1987; Campione, 1989; Campione & Brown, 1987; Spector, 1992).

In terms of research questions and methodological features, some DA studies focus on the amount of learning that accrues on the DA task as a function of student characteristics or the structure of DA (e.g., Tzuriel & Feuerstein, 1992). This approach dominated DA research in the 1970s, 1980s, and 1990s. Alternatively, studies consider DA's contribution in explaining academic performance outside the DA. This second class of studies can be categorized further in terms of two major methodological features. The first is whether studies account for competing predictors of outcome (including static assessments) while considering DA's contribution in predicting academic performance. The second methodological feature is whether academic performance (the outcome) is assessed concurrently with DA (the predictor) or at a later time.

Compared to studies that do not control for competing predictors of outcome, studies that exert such control impose a more stringent falsifiability criterion for considering the value of DA. In terms of the timing of the academic outcome, studies that measure academic performance at a later time enhance external validity, given that the purpose of DA is to forecast future academic achievement. After all, if we were interested in present academic performance, the parsimonious approach would be to measure present achievement directly, not via DA. However, it is also possible that forecasting later academic performance creates variance for DA to capture because the relation between learning potential as indexed via DA and upcoming learning in response to instruction may be stronger than the relation between static assessments (which may be determined by culture, socioeconomics, and previous learning opportunity) and upcoming learning. This empirical issue is, of course, central to questions about DA's utility as a measure of learning potential. For these reasons, in the present study, we examined the contribution of DA in explaining academic performance while accounting for competing predictors of outcome.

To contextualize the present study, we therefore restricted attention in our overview of prior work to the subset of investigations that also explored DA's contribution in predicting academic performance while controlling for competing predictors. We did, however, consider studies that predicted concurrent as well as future academic performance. We also included DAs of varying structure and design. That is, if a study predicted academic outcome while accounting for competing predictors of outcome, we included it regardless of whether the outcome was measured concurrently or in the future, regardless of the way in which DA performance was quantified, regardless of the style of the DA interaction, and regardless of the nature of the DA skills assessed. (For a comprehensive DA review, inclusive of all structures and designs as well as all research questions and methodological features, see Grigorenko & Sternberg, 1998.)



Using our inclusion criteria, we identified three relevant studies.<sup>1</sup> In the first study, Speece, Cooper, and Kibler (1990) measured first-grade students' learning potential with a DA task involving a domain-general skill associated with overall cognitive ability: solving matrices that were borrowed from intelligence tests. Using a standardized style of interaction, the researchers indexed learning potential via the number of prompts required during the assisted phase of assessment. Speece et al. assessed the contribution of DA over verbal IQ, pre-DA matrices performance, and language ability for indicating performance on concurrently administered tests of reading and math achievement. Although statistically significant, DA accounted for less than 2% of the variance in concurrent math performance.

Also predicting concurrent academic performance and using a standardized form of DA, Swanson and Howard (2005) extended the work of Speece et al. (1990) by centering DA on cognitive abilities presumed to underlie reading and math performance: phonological working memory (i.e., rhyming tasks that required recall of acoustically similar words) and semantic working memory (i.e., digit/sentence tasks that required recall of numerical information embedded in short sentences). Four standardized hints were available to DA testers, who selected hints that corresponded to students' errors, choosing the least obvious, relevant hint. Three DA scores were generated: gain score (highest score obtained with assistance), maintenance score (stability of the highest level obtained with assistance probing after assistance was removed), and probe score (number of hints to achieve highest level). The sample comprised students classified as poor readers, skilled readers, reading disabled, or math and reading disabled, with the age of participants averaging 10–12 years. To predict concurrent performance on the Wide Range Achievement Test—Reading and Arithmetic subtests (Wilkinson, 1993), the DA scores for phonological working memory were combined into a factor score; the same was done for DA semantic working memory. The competing predictors were verbal IQ and pre-DA working memory, which were entered first into multiple regression analyses. In predicting reading, pre-DA semantic working memory, verbal IQ, and semantic DA provided unique variance; the unique contribution of the semantic DA factor was 6%. In predicting arithmetic, pre-DA phonological working memory, verbal IQ, and semantic DA provided unique variance; the unique contribution of the semantic DA factor was 25%.

So whereas Speece et al. (1990) found more limited support for DA's added value as related to math performance when DA addressed a domain-general task associated with cognitive ability, Swanson and Howard's (2005) work, which centered DA on cognitive abilities presumed to underlie reading and math performance, was more encouraging. Neither study, however, assessed students' academic performance at a later time. Given that DA's purpose is to forecast future academic achievement, delaying outcome assessment seems important for external validity. In addition, delaying outcome assessment may create variance for DA to capture, as already discussed.

We identified only one study that centered DA on cognitive abilities presumed to underlie reading and math performance (as done by Swanson & Howard, 2005) and, in contrast to Speece et al. (1990) and Swanson and Howard, delayed the assessment of academic achievement to later in the school year. Spector (1992) administered a standardized DA of phonemic awareness, indexing

number of prompts to achieve criterion performance, in November of kindergarten, and then assessed word-level reading skill in May of kindergarten. DA substantially enhanced predictive validity beyond initial verbal ability and beyond initial, static phonological awareness performance in predicting end-of-kindergarten word-reading skill, explaining an additional 21% variance. In fact, November DA was the only significant predictor of May word-reading skill. These results provide promising evidence that DA may enhance the prediction of student learning.

We also considered one additional investigation, even though it did not meet our inclusion criteria (in that it considered DA's relation to post-DA performance on items that mirrored the DA tasks). This study nonetheless helped inform the present study because of its statistical methods. With 84 preschool children, Day, Engelhardt, Maxwell, and Bolig (1997) created DAs that focused on domain-general skills associated with cognitive abilities: similarities and block design tasks. The style of interaction was individualized (i.e., the tester addressed the specific obstacles revealed by the student's responses), and learning potential was indexed by trials to criterion performance. For block design, for example, children were taught a four-step strategy, which the experimenter initially modeled using a four-block problem. If the child failed to solve another four-block problem correctly, more guidance was provided on a three-block problem. Assistance was provided whenever a child failed to solve a problem correctly until he/she solved three consecutive four-block problems without assistance. A transfer task that varied the block design was also presented, with analogous support to learn, and the researchers assessed pre-DA and post-DA similarities and block design performance. Pre-DA performance was considered a competing predictor; post-DA performance was the outcome.

We were interested in Day et al.'s (1997) use of structural equation modeling to test competing measurement and structural models. The measurement model that retained DA as a construct separate from pre-DA and post-DA provided better fit of the data. Thus, DA appeared to tap a construct separate from pre-DA performance, even on analogous tasks. In predicting post-DA scores, Day et al. contrasted four structural models and thereby showed that DA was a viable and necessary predictor of post-DA performance. Of course, academic (or preacademic) performance or learning, which is the relevant outcome to consider when assessing DA's value in predicting achievement, was not considered. Accordingly, Day et al. concluded,

We do not know how dynamic measures related to everyday or school learning or how schooling might affect the relationships among the same measures in children older than those who participated in the present study. . . . Finally, the primary advantage of dynamic measures may lie in how well they predict the ease with which children acquire new information (i.e., training responsiveness) rather than in how well they predict post-training independent performance. (p. 367)

<sup>1</sup> We excluded Byrne's work (e.g., Byrne, Fielding-Barnsley, & Ashley, 2000) because it conceptualizes DA as the student's rate of acquisition in response to schooling; it is not an assessment conducted to predict responsiveness to schooling. Therefore, as an assessment paradigm, Byrne's work is more similar to responsiveness-to-intervention than to DA.



### Purpose of the Present Study

As reflected in Day et al.'s (1997) concluding comments, additional work on DA is necessary, especially in light of Spector's (1992) promising findings. In the present study, the academic outcome was third-grade skill with word problems. Our DA extended the framework for considering type of DA task by employing actual math content that was not a precursor or foundational skill for solving word problems: three basic algebra skills. The question was whether a student's potential to learn algebra at the beginning of third grade forecasted their development of word-problem competence over the course of third grade. We note that although linguistic content is absent from the algebra skills assessed on the DA, algebra does require understanding of the relations among quantities, as is the case for solving word problems. Hence, a potentially important connection does exist between algebraic learning, as assessed on the DA task, and word-problem learning, as addressed in the 16-week treatment conditions (validated vs. conventional word-problems instruction).

We selected basic algebra skills as our DA content for the following reasons. First, we could safely assume that these skills were unfamiliar to third graders (i.e., the skills had not been introduced in school to students by the beginning of third grade and were not familiar from everyday life experiences). Second, the algebra skills were of sufficient difficulty that most third graders would not be able to solve the problems without assistance but could learn the skills with varying amounts of teaching. Third and relatedly, at the beginning of third grade, students should have mastered the simple calculation skills that we incorporated within the three algebra items. Fourth, we could delineate rules underlying the algebra skills, rules that could be used to construct clear explanations within a graduated sequence of prompts. Fifth, the graduated sequences of prompts for the three skills could be constructed in an analogous hierarchy, thereby promoting equal interval scaling of the DA scoring system. Sixth, the three skills were increasingly difficult (as established in pilot work) and later skills appeared to build on earlier skills; therefore, we hoped that transfer across the three algebra skills might facilitate better DA scores. Seventh, as noted, although linguistic content is absent from the algebra skills, algebra does require understanding of the relations among quantities, as is the case for solving word problems.

In the present study, we controlled for other variables that might be important in predicting outcome. First, we controlled for pretreatment performance on salient cognitive predictors of word-problem skill (language ability, attentive behavior, and nonverbal reasoning). Second, we controlled for students' pretreatment skill with calculations and word problems. Third, we extended prior work by also controlling for the nature of classroom instruction. Toward that end, on the basis of random assignment, students received conventional mathematical problem-solving instruction or research-validated schema-broadening instruction (e.g., Fuchs et al., *in press*), and we treated treatment condition as a competing predictor variable (see Method section for a description of these treatment conditions). Two other features of the present study are noteworthy. We examined word-problem outcome as a function of whether the word-problem measures could be considered near versus far transfer from instruction (i.e., the flexibility required to apply the content of the math word-problems curriculum). Also,

we relied on structural equation modeling, as did Day et al. (1997), to assess DA's added value. In the remaining section of this introduction, we briefly explain the basis for selecting language ability, attentive behavior, and nonverbal reasoning as the cognitive predictors of word-problem skill.

Prior work examining cognitive processes that underlie skill with word problems has recurrently identified three important dimensions: attentive behavior, nonverbal reasoning, and language ability. In studies involving attentive behavior, most work has focused on the inhibition of irrelevant stimuli, with mixed results. Passolunghi and colleagues ran a series of studies suggesting the importance of inhibition. For example, comparing good and poor problem solvers, Passolunghi, Cornoldi, and De Liberto (1999) found comparable storage capacity, with inefficiencies of inhibition (i.e., poor problem solvers remembered less relevant but more irrelevant information in math problems). In addition, Fuchs et al. (2005) and Fuchs et al. (2006) studied the role of attention more broadly and, in separate studies, found that a teacher rating scale of attentive behavior predicted the development of skill with word problems at first and third grades.

Nonverbal problem solving, or the ability to complete patterns presented visually, has also been identified as a unique predictor in the development of skill with word problems across first grade (Fuchs et al., 2005), a finding corroborated by Agness and McClone (1987). This is not surprising because word problems, where the problem narrative poses a question entailing relationships between numbers, appear to require conceptual representations, and the finding has been replicated as a unique predictor of concurrent word-problem skill at third grade (Fuchs et al., 2006).

Language ability also is important to consider given the obvious need to process linguistic information when building a problem representation of a word problem. Jordan, Levine, and Huttenlocher (1995) documented the importance of language ability when they showed that kindergarten and first-grade language-impaired children performed significantly lower than nonimpaired peers on word problems. Fuchs, Fuchs, Stuebing, Fletcher, Hamlett, and Lambert (2008) examined concurrent performance on nine cognitive dimensions with multiple measures of calculations skill and word-problem skill on a sample of 917 third graders representatively sampled from 89 classes. Students were classified as having difficulty with calculations, word problems, both domains, or neither domain. Multivariate profile analysis on the nine cognitive dimensions showed that specific calculations difficulty was associated with strength in language, whereas difficulty with word problems was associated with deficient language. We also note that language ability and nonverbal reasoning compose two major dimensions in some prominent assessments of intelligence.

Because these cognitive dimensions are established predictors of word-problem skill, they represent worthy competitors against DA for capturing variance in word-problem outcomes. In addition, these cognitive dimensions, while differing from DA in their demands, connect in transparent ways to word-problem skill or to learning in the general education context. Language, while not reflected in our algebra DA, is clearly involved in word problems, which are communicated via narratives. Nonverbal problem solving asks students to complete matrices presented visually, involving classification and analogy. Although this has no direct connection to our algebra DA, it is linked to word-problem skill, which requires students to connect novel problems with the word-



problem types for which they know solutions. Evidence suggests this occurs via classification and analogy (e.g., Cooper & Sweller, 1987). Attentive behavior asks teachers to judge students' ability to attend to detail, sustain attention, listen, follow directions, organize tasks, keep track of things, ignore extraneous stimuli, and remember daily activities. These ratings, which teachers formulate on the basis of observations of students in their classrooms, seem better connected to learning word problems in general education settings than in one-to-one testing situations like DA, where the tester can redirect and control inattentive behavior more effectively. Finally, we note that language and nonverbal reasoning are two major dimensions of general intelligence, which traditionally has been used to predict academic achievement (in fact, in the present study, the Vocabulary and Matrix Reasoning measures used to index language ability and nonverbal problem solving, respectively, constitute the two-subtest Wechsler Abbreviated Scale of Intelligence [WASI]; Psychological Corporation, 1999). For this additional reason, these constructs represent worthy competitors with DA in forecasting responsiveness to classroom word-problem instruction.

## Method

### Participants

With the exception of the DA, the data described in this article were collected as part of a prospective 4-year study assessing the effects of mathematical problem-solving instruction and examining the developmental course and cognitive predictors of mathematical problem solving. The data in the present article were collected with a subset of students in the 4th-year cohort of the larger study at the first and second assessment waves (at fall and spring of third grade). We sampled students from the 30 participating classrooms in nine schools (five Title I and three non-Title I). Classrooms had been randomly assigned to treatment conditions (conventional vs. validated schema-broadening instruction) within schools. There were 2–6 students per class.<sup>2</sup>

We sampled students for the present study from the cohort's 510 students with parental consent. The sampling process was designed to yield a representative sample. That is, from these 510 students, we randomly sampled 150 for participation, blocking within instructional condition (conventional math problem-solving instruction vs. schema-broadening math problem-solving instruction), within classroom, and within three strata: (a) 25% of students with scores 1 *SD* below the mean of the entire distribution on the Test of Computational Fluency (Fuchs, Hamlett, & Fuchs, 1990; see *Measures of Cognitive, DA, and Calculations Pretreatment Performance*), (b) 50% of students with scores within 1 *SD* of the mean of the entire distribution on the Test of Computational Fluency, and (c) 25% of students with scores 1 *SD* above the mean of the entire distribution on the Test of Computational Fluency. Of these 150 students, we have complete data for the variables reported in the present study on 122 children. The 122 included students were comparable to the remaining students on the study variables, with varying sample sizes depending on where missing data occurred. See Table 1 for descriptive information on performance variables for the sample of 122 children. Of these students, 67 (54.9%) were male, and 80 (67.0%) received subsidized lunch. Ethnicity was distributed as follows: 57 (47.5%) African Ameri-

can, 53 (43.4%) European American, 10 (8.3%) Hispanic, and 3 (2.5%) other. Two students (1.7%) were English language learners.

On the basis of random assignment, 61 students received conventional math problem-solving instruction (as determined by their core math program and teachers), and 61 received schema-broadening math problem-solving instruction that has been validated (i.e., five large-scale randomized control trials, published in peer-reviewed journals, have documented the efficacy of the intervention at third grade with similar populations of students; Fuchs et al., 2003a, 2003b; Fuchs, Fuchs, Finelli, et al., 2004; Fuchs, Fuchs, Prentice, et al., 2004; Fuchs et al., in press). In Table 1, we show performance variables by instructional condition. Of the 61 students in conventional math problem-solving instruction, 32 (52.5%) were male, and 39 (63.9%) received subsidized lunch. Ethnicity was distributed as follows: 32 (52.5%) African American, 25 (41.0%) European American, and 5 (8.2%) Hispanic. Two students (3.2%) were English language learners. Of the 61 students in validated, schema-broadening math problem-solving instruction, 35 (57.4%) were male, and 41 (67.2%) received subsidized lunch. Ethnicity was distributed as follows: 25 (41.1%) African American, 28 (45.9%) European American, 5 (8.2%) Hispanic, and 3 (4.9%) other. None was an English language learner.

### Procedure

We describe the subset of measures relevant to this research report. In October, the DA was administered individually in one 30- to 45-min session. In September and October, we administered measures of language and nonverbal reasoning (Woodcock Diagnostic Reading Battery [WDRB]—Listening Comprehension; Woodcock, 1997; Test of Language Development—Primary [TOLD] Grammatical Closure; Newcomer & Hammill, 1988; WASI Vocabulary; and WASI Matrix Reasoning) and math performance (Woodcock-Johnson III Tests of Cognitive Abilities [WJ III] Applied Problems; Woodcock, McGrew, & Mather, 2001) individually in two 45-min sessions. In October, we administered the Test of Algorithmic Word Problems (Fuchs et al., 2003a) in one 30-min large-group session, and we administered three tests of calculations skill (Addition Fact Fluency [Fuchs, Hamlett, & Powell, 2003], Subtraction Fact Fluency [Fuchs, Hamlett, & Powell, 2003], and Test of Mixed Algorithms [Fuchs et al., 1990]) in one 60-min large-group session. We also obtained scores on the previous spring's state assessment (Tennessee Comprehensive Assessment Program; CTB/McGraw-Hill, 2003) from teachers.

In March, five tests of word-problem skill were administered (Algorithmic Word Problems, Complex Word Problems, Real-World Problem Solving [all three from Fuchs et al., 2003a], and

<sup>2</sup> Concerning classroom as a potential source of dependency in the data, we calculated intraclass correlations (ICCs) for classrooms on the outcomes measures, which is where important context effects might be revealed. We found minimal ICCs for the far-transfer measures (.003–.03). By contrast, the ICCs for the near-transfer measures were sizeable (.5–.7). However, the ICCs for the near-transfer measures dropped to levels comparable to the far-transfer measures when the effects of treatment were controlled, as in the SEM models. So in the SEM models, variance in the outcomes is essentially at the individual not the classroom level, making it unnecessary to account for classroom in the SEM models.



Table 1  
Performance For Total Sample and by Treatment Condition

Variable	Condition											
	Total sample				Conventional				SBI			
	Raw score		Standard score		Raw score		Standard score		Raw score		Standard score	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Descriptive												
WASI IQ			96.10	12.55			96.60	13.28			95.58	11.83
TCAP Reading			54.74	14.55			53.81	14.67			55.61	14.51
TCAP Math			60.70	17.90			59.48	18.18			61.87	17.71
WJ Applied Problems	30.04	4.15	106.97	12.58	29.90	4.58	106.52	12.90	30.18	3.68	107.43	12.33
WRMT-WID	56.66	11.51	100.94	11.20	55.84	12.44	100.42	12.19	57.50	10.49	101.14	10.14
Pretreatment performance												
Grammatic Closure	19.67	5.84	86.43	10.47	19.06	6.43	85.56	10.91	20.30	5.14	87.33	10.02
Listening Comprehension	21.64	3.92	98.88	17.47	21.93	4.11	99.32	16.44	21.35	3.59	98.44	18.06
Vocabulary	26.39	5.82	45.11	9.19	26.26	6.26	44.98	9.94	26.52	5.38	45.23	8.42
Matrix Reasoning	16.08	5.95	49.65	10.22	16.35	5.90	50.15	10.14	15.80	6.03	49.13	10.36
SWAN 1-3	12.61	3.89		12.42	4.04		12.80	3.66				
SWAN 4-6	12.70	4.17			13.07	4.47			12.33	3.74		
SWAN 7-9	12.70	4.06			12.98	4.47			12.42	3.52		
DA1	5.25	1.53			5.29	1.69			5.20	1.36		
DA2	3.34	1.81			3.23	1.94			3.45	1.68		
DA3	1.39	1.99			1.35	1.91			1.42	2.06		
Addition Fact Fluency	12.25	4.84			12.64	5.20			11.85	4.45		
Subtraction Fact Fluency	7.84	5.19			8.36	5.27			7.32	5.09		
Test of Mixed Algorithms	23.30	10.23			22.82	10.93			23.58	9.48		
Algorithmic Word Problems	8.80	5.45			8.21	5.32			9.39	5.41		
Outcomes												
Real-World Word Problem Solving	24.38	13.96			21.65	12.97			27.21	14.49		
Iowa	15.74	4.36	188.05	20.60	15.44	4.24	186.61	19.97	16.05	4.50	189.53	21.3
WJ Applied Problems	32.39	4.32	106.97	12.58	32.19	4.64	106.52	12.90	32.60	4.19	107.43	12.33
Algorithmic Word Problems	23.94	12.27			15.69	9.04			15.69	9.04		
Complex Word Problems	27.05	16.86			16.26	10.52			38.20	14.81		

Note. SBI = schema-broadening instruction; WASI = Wechsler Abbreviated Scale of Intelligence; TCAP = Tennessee Comprehensive Assessment Program; WJ = Woodcock-Johnson III Tests of Cognitive Abilities; WRMT-WID = Woodcock Reading Mastery Tests—Revised: Word Identification; SWAN = SWAN Rating Scale; DA = dynamic assessment; Iowa = Iowa Test of Basic Skills: Problem Solving and Data Interpretation.

WJ III Applied Problems). The first four were administered in two sessions in large groups. WJ III Applied Problems was administered individually in one session. We categorized the word-problem measures in terms of transfer distance (i.e., the distance from the types of word problems addressed during instruction and the degree of flexibility required in applying the word problems taught during instruction). Algorithmic Word Problems and Complex Word Problems were deemed near transfer. Real-World Problem Solving, WJ III Applied Problems, and the Iowa Test of Basic Skills: Problem Solving and Data Interpretation (Hoover, Dunbar, & Frisbie, 2001) were considered far transfer. (We elaborate on transfer distance for each measure in the measures section.)

Group and individual testing was conducted by trained university examiners, each of whom had demonstrated 100% accuracy during mock administrations. All individual sessions were audio-taped, and 19.9% of tapes, distributed equally across testers, were selected randomly for accuracy checks by an independent scorer. Agreement was between 98.7% and 99.9%. In October, classroom teachers completed the SWAN Rating Scale (Swanson, 2004), the measure of attentive behavior, on each student.

### Measures of Cognitive, DA, and Calculations Pretreatment Performance

**Language.** We used three measures of language skill to create a latent variable representing language. TOLD Grammatic Closure (Newcomer & Hammill, 1988) measures the ability to recognize, understand, and use English morphological forms. The examiner reads 30 sentences, one at a time; each sentence has a missing word. As per the test developers, for 8-year-olds, reliability is .88; the correlation with the Illinois Test of Psycholinguistic Ability Grammatic Closure (Kirk, McCarthy, & Kirk, 1968) is .88. Coefficient alpha on the representative sample was .76. The WDRB Listening Comprehension (Woodcock, 1997) measures the ability to understand sentences or passages. Students supply words missing from the end of each sentence or passage. The test begins with simple verbal analogies and associations and progresses to comprehension involving the ability to discern implications. Testing is discontinued after six consecutive errors. The score is the number of correct responses. As per the test developers, reliability is .80 at ages 5–18; the correlation with Woodcock-Johnson Psycho-Educational Battery—Revised (Woodcock & Johnson, 1989) is

.73. Coefficient alpha on the representative sample was .81. WASI Vocabulary (Psychological Corporation, 1999) measures expressive vocabulary, verbal knowledge, and foundation of information. The first four items present pictures; the student identifies the object in the picture. For remaining items, the tester says a word that the student defines. Testing is discontinued after five consecutive errors. As reported by Zhu (1999), split-half reliability is .86–.87 at ages 6–7; the correlation with Wechsler Intelligence Scale for Children (3rd rev.; Wechsler, 1991) Full Scale IQ is .72. Coefficient alpha on the representative sample was .78.

*Nonverbal problem reasoning.* WASI Matrix Reasoning (Psychological Corporation, 1999) measures nonverbal reasoning with pattern completion, classification, analogy, and serial reasoning. Examinees look at a matrix from which a section is missing and complete it. Testing is discontinued after four errors on five consecutive items or four consecutive errors. As per the test developer, reliability is .94 for 8-year-olds; the correlation with Wechsler Intelligence Scale for Children (3rd rev.) Full Scale IQ is .66. Coefficient alpha on the representative sample was .76.

*Attentive behavior.* The SWAN Inattentive subscale is a 9-item teacher rating scale. Items from the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; American Psychiatric Association, 1994) criteria for attention-deficit/hyperactivity disorder are included. Items are rated on a 1–7 scale. We grouped items into triads (1–3, 4–6, 7–9) to form three observed attentive behavior variables for use in the structural equation modeling. The SWAN has been shown to correlate well with other dimensional assessments of behavior related to inattention (Swanson, 2004). Coefficient alpha in this study was .92 for items 1–3, .94 for items 4–6, and .94 for items 7–9.

*DA.* The DA addresses three algebra skills assumed to be novel because they have not been taught in school and because they do not routinely occur in children's extra-school experiences. The three algebra skills are (a) finding the missing variable in the first or second position in addition equations (e.g.,  $x + 5 = 11$  or  $6 + x = 10$ ), (b) finding  $x$  in multiplication equations (e.g.,  $3x = 9$ ), and (c) finding the missing variable in equations with two missing variables but with one variable then defined (e.g.,  $x + 2 = y - 1$ ;  $y = 9$ ). We refer to these three skills, respectively, as *DA Skill A*, *DA Skill B*, and *DA Skill C*.

Mastery of each DA skill is assessed before instructional scaffolding begins, and mastery testing recurs after each level of instructional scaffolding is completed. The mastery test comprises six items representing the skill targeted for mastery, with mastery defined as at least five items correct. The items on the test are never used for instruction but parallel instructional items; each time the six-item test is readministered for a given skill, a different form is used, although some items recur across forms. If, after 5 s, the student has not written anything and does not appear to be working, the tester asks, "Can you try this?"; if after another 15 s, the student still has not written anything and does not appear to be working, the tester asks, "Are you still working or are you stuck?" If the student responds that he/she is stuck, the tester initiates the first (or next) level of instructional scaffolding. If the student responds that he/she is still working but another 30 s pass without the student writing anything or working, the tester then initiates the first (or next) level of instructional scaffolding. If the student masters the skill (i.e., at least five items are answered correctly), the tester administers a generalization problem (i.e., for DA Skill

A:  $3 + 6 + x = 11$ ; for DA Skill B:  $14 = 7x$ ; for DA Skill C:  $3 + x = y + y$ ;  $y = 2$ ) and moves to the next DA skill. If the student does not master the skill (i.e., fewer than five items are answered correctly), the tester provides the first (or next) level of instructional scaffolding, or support, which is followed by the six-item test. The levels of instructional scaffolding gradually increase instructional explicitness and concreteness. If a student fails to answer at least five items correctly after the tester provides all five scaffolding levels for a given skill, the DA is terminated.

Scores range from 0 to 21, where 0 is the worst score (i.e., student never masters any of the three skills) and 21 is the best score (i.e., student masters each of the three skills on the pretest and gets every bonus problem correct). So, for each skill, there is a maximum of 7 points, awarded as follows: student masters skill on pretest = 6 points, student masters skill after Scaffolding Level 1 = 5 points; student masters skill after Scaffolding Level 2 = 4 points, student masters skill after Scaffolding Level 3 = 3 points, student masters skill after Scaffolding Level 4 = 2 points, student masters skill after Scaffolding Level 5 = 1 point, student never shows mastery = 0 points. In addition, if the student gets the generalization problem correct, 1 point is added, for the maximum score of 7 points for that DA skill.

To promote equal interval scaling, scaffolding levels for each of the three skills are structured analogously. Scaffolding levels, which range from incidental to explicit, are provided in the Appendix for DA Skill A. For this information on DA Skills B and C, contact Lynn S. Fuchs. Correlations with the previous year's math composite score on the state assessment (CTB/McGraw-Hill, 2003) were .57 for DA Skill A, .60 for DA Skill B, and .41 for DA Skill C.

*Calculations.* Three measures were used to create a latent variable representing calculation skill. Addition Fact Fluency (part of Grade 3 Math Battery; Fuchs, Hamlett, & Powell, 2003) comprises 25 addition fact problems with answers from 0 to 12 and with addends from 0 to 9. Problems are presented horizontally on one page. Students have 1 min to write answers. Percentage of agreement, on 20% of protocols by two independent scorers, was 99.9. Coefficient alpha for this sample was .92; criterion validity with the previous spring's TerraNova (CTB/McGraw-Hill, 2003) Total Math score was .53 for the 1139 students on whom we had TerraNova scores. Subtraction Fact Fluency (part of Grade 3 Math Battery; Fuchs, Hamlett, & Powell, 2003) comprises 25 subtraction fact problems with answers from 0 to 12 and with minuends/subtrahends from 0 to 18. Problems are presented horizontally on one page. Students have 1 min to write answers. Percentage of agreement, on 20% of protocols by two independent scorers, was 98.7. Coefficient alpha for this sample was .93, and criterion validity with the previous spring's TerraNova (CTB/McGraw-Hill, 2003) Total Math score was .51 ( $n = 1139$ ). The Test of Mixed Algorithms (Fuchs, Hamlett, & Fuchs, 1990) is a one-page test with 25 items that sample the typical second-grade computation curriculum, including adding and subtracting number combinations and procedural computation. Students have 3 min to complete as many answers as possible. Staff entered responses into a computerized scoring program on an item-by-item basis, with 15% of tests reentered by an independent scorer. Data-entry agreement was 99.7. Coefficient alpha for this sample was .94, and criterion validity with the previous spring's TerraNova (CTB/McGraw-Hill, 2003) Total Math score was .60 ( $n = 1139$ ).



### Measures of Word-Problem Skill

We measured pretreatment word-problem skill using Algorithmic Word Problems. We measured posttreatment word-problem outcomes using Algorithmic Word Problems, Complex Word Problems, Real-World Problem Solving, WJ III Applied Problems, and Iowa Test of Basic Skills: Problem Solving and Data Interpretation.

Four independent judges classified the five word-problem outcome measures in terms of transfer distance from the instruction provided in schema-broadening math problem-solving instruction (i.e., the flexibility required to apply the content of the math word-problem curriculum) into two classes: near transfer versus far transfer. Agreement was 100%. We describe the word-problem outcomes in terms of near-transfer measures (all novel problems but structured similarly to problems used for instruction) versus far-transfer measures (all novel problems, not structured similarly to problems used for instruction). Schema-broadening instruction (SBI) focused on four word-problem types, chosen from the district curriculum to ensure that conventional math problem-solving instruction students had instruction relevant to the study. The four problem types were “shopping list” problems (e.g., “Joe needs supplies for the science project. He needs 2 batteries, 3 wires, and 1 board. Batteries cost \$4 each, wires cost \$2 each, and boards cost \$6 each. How much money does he need to buy supplies?”), “half” problems (e.g., “Marcy will buy 14 baseball cards. She’ll give her brother half the cards. How many cards will Marcy have?”), step-up function or “buying bags” problems (e.g., “José needs 32 party hats for his party. Party hats come in bags of 4. How many bags of party hats does José need?”), and two-step “pictograph” problems (e.g., “Mary keeps track of the number of chores she does on this chart [pictograph is shown with label: each picture stands for 3 chores]. She also took her grandmother to the market 3 times last week. How many chores has Mary done?”). In addition, SBI was designed to broaden schemas, with each problem varying the cover story and one of the four transfer features, which change a problem without altering its type or solution: a problem with unfamiliar vocabulary, posing an additional question, incorporating irrelevant information, or combining problem types.

*Near-transfer word problems.* Algorithmic Word Problems (Fuchs et al., 2003a) comprises 10 word problems each of which requires 1–4 steps. The measure samples the four problem types that correspond to SBI and that are incorporated within conventional classroom instruction: shopping list, half, buying bags, and pictograph problems. The tester reads each item aloud while students follow along on their own copies of the problems; the tester progresses to the next problem when all but 1–2 students have their pencils down, indicating they are finished. Students can ask for rereading(s) as needed. The maximum score is 44. We used two alternate forms; the problems in both forms required the same operations, incorporated the same numbers, and presented text with the same number/length of words. In half the classes in each treatment condition, we used Form A at pretest and Form B at posttest; in the other half, forms were reversed. For the representative sample, Cronbach’s alpha was .85, and criterion validity with the previous spring’s TerraNova (CTB/McGraw-Hill, 1997) Total Math score was .58 for the 844 students on whom we had TerraNova scores. Interscorer agreement, computed on 20% of protocols by two independent scorers, was .984.

Complex Word Problems (Fuchs et al., 2003a) comprises nine problems representing the same four problem types within more complex contexts, which incorporate the four transfer features explicitly taught within SBI: (a) adding multiple quantities of items with different prices, with information presented in bulleted format and with a selection response format; (b) adding multiple quantities of items with different prices, also asking for money left at the end; (c) a step-up function problem with irrelevant information; (d) a step-up function that requires students to compare the prices of two packaging options; (e) a half problem using the words *share equally* instead of *half*; (f) a pictograph/adding problem asking for money left at the end; (g) a pictograph/adding problem comparing two quantities; (h) a problem with irrelevant information that combined multiple quantities with different prices and pictograph/adding; and (i) a problem with irrelevant information that combined multiple quantities with different prices and a step-up function. The tester reads each item aloud while students follow along on their own copies of the problems; the tester progresses to the next item when all but 1–2 students have their pencils down, indicating they are finished. Students can ask for rereading(s) as needed. The maximum score is 79. For the representative sample, Cronbach’s alpha was .88, and criterion validity with the previous spring’s TerraNova Total Math Score was .55 for the 844 students on whom we had TerraNova scores. Interscorer agreement, computed on 20% of protocols by two independent “blind” scorers, was .983.

*Far-transfer word problems.* Real-World Problem Solving (Fuchs et al., 2003a) simultaneously assesses transfer of all four problem types and all four transfer features addressed in SBI. Also, to decrease association between the task and classroom or tutoring SBI, far transfer was formatted to look like a commercial test (printed with a formal cover, on green paper, with photographs and graphics interspersed throughout the test booklet). Two assessments were constructed as alternate forms: Although the context of the problem situations differed, the structure of the problem situation and the questions were identical, and the problem solutions and reading demands were equivalent.

Performance was scored according to a rubric with four dimensions: conceptual underpinnings, computational applications, problem-solving strategies, and communicative value. The original rubric (Kansas State Board of Education, 1991) scored responses on a 6-point scale. To enhance reliability, we awarded points on a finer basis (e.g., the problem-solving strategies score included points for finding relevant information, accumulating to a total, showing all computation, working the answer in distinct multiple parts, labeling at least half of the multiple parts, and labeling work with monetary and operation signs). Across the four questions and four scoring dimensions, the maximum score is 72. On this sample, Cronbach’s alpha was .91 at pretest; .94 at posttest; concurrent validity with WJ III Applied Problems (Woodcock et al., 2001) was .47 at pretest and .61 at posttest. Interscorer agreement, computed on 20% of protocols by two independent “blind” scorers, was .987 at pretest and .949 at posttest. Given the deleterious effects of student unfamiliarity with performance assessments (Fuchs, Fuchs, Karns, et al., 2000), research assistants delivered a 45-min “test-wiseness” lesson before pre- and posttesting in all conditions. (The mean score across immediate, near, and far transfer correlated .62 with WJ III Applied Problems [Woodcock et al., 2001] at pretest and .57 at posttest.)



WJ III Applied Problems (Woodcock et al., 2001) measures skill in analyzing and solving practical math problems with 60 items. The tester orally presents items involving counting, telling time or temperature, and problem solving. The word problems addressed a range of problem types, none of which was directly addressed within SBI. Testing is discontinued after six consecutive errors. The score is the number of correct items. As reported by McGrew and Woodcock (2001), 1-year test-retest reliability is .85; the ratio of true score variance to observed variance is .88-.91. Coefficient alpha on this sample was .85.

*Iowa Test of Basic Skills: Problem Solving and Data Interpretation.* With Iowa Test of Basic Skills: Problem Solving and Data Interpretation (Hoover et al., 2001), students solve 24 word problems and use data presented in tables and graphs to solve word problems. The problems on this test were related only marginally to the content of SBI. Five problems included pictographs embedded within problem types not addressed in SBI. One problem incorporated irrelevant information, again embedded in a problem type not addressed in SBI. For Grades 1-5, Kuder-Richardson Formula 20 is .83-.87. In this study, coefficient alpha was .86.

### Instruction

On the basis of random assignment, students received conventional math problem-solving or validated SBI. We describe these conditions briefly. For additional information, see Fuchs et al. (in press). The problem types were already described under *Measures of Word-Problem Skill*.

*Conventional instruction.* To guide instruction relevant to the four problem types, conventional instruction relied primarily on Houghton Mifflin Math (Greenes et al., 2007). Instruction addressed one problem type at a time (as did SBI) and focused on the concepts underlying the problem type. In addition, a prescribed set of problem-solution rules was taught, with explicit steps for arriving at solutions to the problems presented in the narrative. There was no attempt to broaden students' schemas for these problem types to address transfer. However, in comparison to the classroom SBI, classroom control group instruction provided more practice in applying problem-solution rules and provided greater emphasis on computational requirements. Conventional instruction was explicit and relied on worked examples, guided group practice, independent work with checking, and homework. In addition, conventional instruction (as well as SBI) incorporated a 3-week researcher-designed and delivered general problem-solving strategies unit.

*General problem-solving strategies instruction (conventional instruction and SBI).* Conventional instruction and SBI students received a researcher-designed 3-week (2 lessons per week) instructional unit on general math problem-solving strategies, which was conceptually unrelated to SBI. It addressed making sure answers make sense; lining up numbers from text to perform math operations; checking computation; and labeling work with words, monetary signs, and mathematical symbols. These six lessons, each lasting 30-40 min, relied on worked examples with explicit instruction, dyadic practice, independent work with checking, and homework, for a total of 210 min. For a manual with general problem-solving strategies, contact Lynn S. Fuchs.

*SBI.* For an SBI manual, contact Lynn S. Fuchs. SBI students received the 3-week general math problem-solving unit as well as four researcher-designed 3-week SBI units. Each SBI unit com-

prised six sessions. Also, two cumulative review sessions were delivered the week after winter break. In each unit, Sessions 1 and 5 lasted about 40 min; the others lasted about 30 min. This totaled 200 min per unit and 856 min across the units (including the two cumulative review sessions). Each 3-week unit addressed one of the four problem types: shopping list, buying bag, half, and pictograph.

Within each unit, the sequence of lessons was as follows. In Sessions 1-4, problem-solution instruction was delivered, using problems that varied only cover stories. A poster listing the steps of the solution method was displayed in the classroom. In Session 1, teachers addressed the underlying concepts and structural features for the problem type, presented a worked example, and, as they referred to the poster, explained how each step of the solution method had been applied in the example. Students responded frequently to questions. After reviewing the concepts and presenting several worked examples in this way, teachers shared partially worked examples while students applied the solution steps. Students then completed 1-4 problems in dyads, where stronger students helped weaker students solve problems and check work with answer keys. Sessions 2-4 were structured similarly, with a greater proportion of time spent on partially worked examples and dyadic practice. Also, at the end of Sessions 2-4, students completed one problem independently, the teacher checked work against an answer key, and students graphed scores.

Sessions 5-6 were designed to broaden schemas, with each problem varying the cover story and one of the four transfer features addressed in SBI. Teachers first taught the meaning of the word *transfer*. Then, they taught the four transfer features, which change a problem without altering its type or solution: A familiar problem type, for which a solution is known, can use unfamiliar vocabulary, can pose an additional question, can incorporate irrelevant information, can combine problem types. A poster, "Transfer: Ways Problems Change," was displayed. In Session 5, teachers explained the poster, illustrating each transfer feature with a worked example. They gradually moved to partially worked examples. Then, students worked in pairs to apply the solution method to problems that varied transfer features. In Session 6, teachers reviewed the four transfer features using similar procedures, except students spent more time working in dyads and then completed a problem independently, scored work against a key, and graphed scores.

*Delivery.* Each research assistant, all full-time research employees or graduate students in the College of Education, had responsibility for students in both conditions. All sessions were scripted to ensure consistency of information; however, to permit natural teaching styles, scripts were studied, not read. To ensure comparable mathematics instructional time across conditions, SBI sessions occurred within the confines of the mathematics instructional block. At the end of the study, classroom teachers reported the number of minutes per week they spent on math, including time on SBI, and the amount of instructional time for conventional versus SBI students was comparable and similar.

*Treatment fidelity.* Prior to the first delivery of each session, research assistants agreed on the essential information in the script and made a checklist of points. Each session was audiotaped. At the study's end, two research assistants independently listened to tapes while completing the checklist to identify the percentage of points addressed. We sampled tapes so that, within conditions,



research assistants and lesson types were sampled equitably. In conventional instruction, 1–2 tapes were sampled per class (for Unit 1); in SBI, 6–7 tapes were sampled per class (distributed equally across Units 1–5). Intercoder agreement, calculated on 20% of the sampled tapes, was 97.2%. The mean percentage of points addressed was 98.34 ( $SD = 3.16$ ) for conventional instruction and 95.82 ( $SD = 2.46$ ) for SBI.

## Data Analysis and Results

### Descriptive Data

Data analysis progressed in two stages. First, the measurement models for pretreatment and posttreatment measures were estimated using confirmatory factor analysis. Second, various relational models were tested using structural equation modeling. Prior to model estimation, the data were preanalyzed to identify outliers and estimate departures from univariate and multivariate normality. Univariate plots revealed no significant outliers (plus or minus three standard deviations from the mean of the sample on that variable). PRELIS 8.7 (Jöreskog & Sörbom, 2004) was used to explore univariate and multivariate normality. In the case of univariate normality, PRELIS provides separate estimates of skew and kurtosis for each variable with accompanying  $p$  values. Several variables exhibited significant skew (Complex Word Problems, Subtraction Fact Fluency, and DA Skill A) and kurtosis (Real-Word Problem Solving and Subtraction Fact Fluency). Mardia's statistic for multivariate normality revealed significant multivariate nonnormality for skew ( $z = 2.361$ ,  $p = .018$ ) but not for kurtosis ( $z = 0.787$ ,  $p = .431$ ). To examine the extent to which nonnormality affects the chi-square fit statistic, models were constructed using maximum likelihood versus a scaled chi-square estimated with robust standard errors using the method developed by Satorra and Bentler (1994) as suggested by West, Finch, and Curran (1994). A scaling correction factor was calculated representing the standard chi-square divided by the scaled chi-square. Scaling correction factors ranged from 0.97 to 1.05 across models suggesting little difference between the standard and scaled chi-square values. Thus, we concluded that multivariate nonnormality had little effect on the estimated chi-square in this study. For the remainder of the article, we report results from models estimated with LISREL 8.7 (Jöreskog & Sörbom, 2004) using a maximum likelihood estimation on the sample of participants with complete data. There is growing evidence to suggest that maximum likelihood-based fit indices outperform other estimation procedures and perform reasonably well with small sample sizes even when distributions are less than optimal (see Hu & Bentler, 1998).

In Table 1, we show means and standard deviations to describe the sample on pretreatment performance variables and on the math outcome variables. We provide this for the total sample and by instructional condition. In Table 2, we show correlations among the pretreatment performance and the math outcome study variables for the total sample. We note that all observed (manifest) variables were transformed to  $z$  scores prior to structural equation modeling. Also in Table 2, we show correlations between treatment condition (validated SBI vs. conventional) and the math outcome study variables (we do not provide the correlations between treatment condition and pretreatment performance variables because treatment was randomized).

### *Pretreatment Measurement Model: Is DA a Distinct Dimension of Pretreatment Ability?*

Our pretreatment measurement model included six dimensions of pretreatment performance. The first latent variable, language ability, comprised TOLD Grammatical Closure, WDRB Listening Comprehension, and WASI Vocabulary. The second dimension, attentive behavior, included three clusters of items from the SWAN (Items 1–3, Items 4–6, and Items 7–9). The third latent variable, DA, comprised DA Skill A, DA Skill B, and DA Skill C. The fourth latent variable, calculations skill, incorporated Addition Fact Fluency, Subtraction Fact Fluency, and Test of Mixed Algorithms. The fifth latent variable, a single manifest variable, WASI Matrix Reasoning, was used to measure nonverbal reasoning. Within structural equation modeling, the default is to set the error variance to zero for manifest variables. Instead of using the default, we accounted for measurement error in nonverbal reasoning using the alpha coefficient for the present sample ( $r = .94$ ) to specify the error variance. A reliability of .94 is equivalent to an error variance of 0.06 times the variance of nonverbal reasoning, which in this case was 1.0 (for details, see MacCallum, 1995). The final latent variable, Algorithmic Word Problems, was also a single manifest variable. It was used to measure pretreatment word-problem skill. To account for measurement error in word-problem skill, we used the alpha coefficient for the present sample ( $r = .84$ ) to specify the error variance of 0.16.

See Table 3 for correlations among the pretreatment measurement model latent traits. All observed variables loaded substantially and reliably onto their respective factors (standardized coefficients: .44 to .98,  $ps < .001$ ). Because this base measurement model included the greatest number of estimated parameters (compared to the competing measurement models), it provided the best characterization of the data and served as a basis for comparing the competing, more parsimonious models. Chi-square goodness-of-fit statistic plus four different fit indices (representing Type 1–3 fit and absolute-fit indices) are provided. Jaccard and Wan (1996) and Kline (1998) suggested use of multiple fit indices to reflect diverse criteria. Hu and Bentler (1998) recommended with maximum likelihood estimation methods that standardized root-mean-square residual (SRMR; absolute fit) be reported and supplemented with normed fit index (NFI; Type 1 fit), nonnormed fit index (NNFI; Type 2 fit), and comparative fit index (CFI; Type 3 fit). In evaluating a model, adequate fit is indicated by a nonsignificant chi-square test; NFI, NNFI, and CFI exceeding .95; and a SRMR of less than .08 (see Hu & Bentler, 1998). The base six-factor measurement model accounted well for the data structure,  $\chi^2(64, N = 122) = 58.31$ ,  $p = .6771$ ; NFI = .965; NNFI = 1.000; CFI = 1.000; and SRMR = 0.0389 (see Pretreatment Measurement Model 1 in Table 4 and Figure 1).

To assess the distinctiveness of DA as a pretreatment performance dimension, we compared this base measurement model against five more parsimonious, five-factor measurement models. The utility of one model to explain data can be compared statistically against the utility of other models in which it is nested by using a chi-square difference tests (i.e.,  $\Delta\chi^2$ ). The base model always yields the best fit because it includes the most estimated parameters and is therefore the least restrictive model. However, in the interest of parsimony, a more restrictive nested model that fails to yield a significantly worse fit is accepted as superior.

Table 2  
Correlations Among Study Variables

Variable	Treat <sup>a</sup>	Language					Attention			Dynamic Assessment			Calculations			Pretreatment word problems	Word problems: Far transfer			Word problems: Near transfer		
		GC	LC	V	MR	S13	S56	S79	DA1	DA2	DA3	A	S	M	AL1	RW	I	AP	AL2	C		
Pretreatment Performance Language																						
Grammatic Closure (GC)	b	—																				
Listening Comprehension (LC)	b	.59	—																			
Vocabulary (V)	b	.55	.57	—																		
Nonverbal Reasoning	b	.17	.24	.23	—																	
Matrix Reasoning (MR)																						
Attention																						
SWAN 1–3 (S13)	b	.36	.33	.42	.21	—																
SWAN 4–6 (S46)	b	.36	.30	.38	.19	.93	—															
SWAN 7–9 (S79)	b	.34	.29	.37	.10	.90	.93	—														
Dynamic Assessment (DA)																						
DA1	b	.30	.30	.29	.50	.36	.33	.29	—													
DA2	b	.36	.37	.35	.44	.43	.43	.37	.58	—												
DA3	b	.20	.27	.33	.29	.20	.15	.07	.28	.35	—											
Calculations																						
Addition Fact Fluency (A)	b	.14	.17	.25	.29	.41	.43	.37	.46	.37	.17	—										
Subtraction Fact Fluency (S)	b	.18	.17	.25	.28	.39	.42	.35	.39	.42	.28	.73	—									
Mixed Algorithms (M)	b	.17	.17	.20	.24	.43	.47	.40	.40	.47	.24	.62	.64	—								
Word Problems																						
Algorithmic Word Problems (AL1)	b	.36	.39	.34	.35	.43	.41	.34	.42	.51	.25	.34	.37	.32	—							
Outcomes																						
Word Problems—Far Transfer																						
Real-World Word Problems (RW)																						
Iowa (I)	.20	.38	.45	.48	.41	.53	.50	.42	.41	.58	.34	.40	.43	.44	.67	—						
WJ Applied Problems (AP)	.07	.44	.50	.44	.38	.50	.46	.42	.48	.52	.36	.34	.42	.36	.64	.66	—					
Word Problems—Near Transfer	.05	.39	.45	.46	.46	.45	.49	.42	.60	.68	.36	.52	.52	.57	.61	.66	.67	—				
Algorithmic Word Problems (AL2)																						
Complex Word Problems (C)	.69	.23	.26	.19	.14	.33	.36	.32	.24	.30	.09	.22	.17	.31	.43	.49	.40	.39	—			
	.65	.25	.26	.22	.11	.37	.37	.33	.23	.29	.07	.24	.15	.28	.45	.54	.41	.39	.82	—		

Note. WJ = Woodcock-Johnson III Tests of Cognitive Abilities.

<sup>a</sup> Treat = treatment condition, where 1 = conventional instruction on word problems and 2 = validated SBI on word problems. <sup>b</sup> The correlation between treatment condition and the dimensions of pretreatment performance were set to zero, given random assignment to treatment.



Table 3  
Latent Trait Correlations

Variable	L	NVR	A	DA	WP
Pretreatment					
Language (L)	—				
Nonverbal reasoning (NVR)	.31	—			
Attention (A)	.48	.18	—		
Dynamic assessment (DA)	.60	.64	.50	—	
Calculations (C)	.31	.33	.52	.67	—
Word problems (WP)	.52	.40	.45	.68	.46

Note. The correlation between the near- and far-transfer outcome latent variables was .49.

In each of the five competing five-factor measurement models, we merged DA with one other dimension of pretreatment performance. So, for example Pretreatment Measurement Model 2 (see Table 4) included a separate attentive behavior factor, a separate calculations skill factor, a separate nonverbal reasoning factor, pretreatment word-problem skill, and a factor in which indices of language and DA loaded together (on a single factor). In Pretreatment Measurement Models 3, 4, 5, and 6 (see Table 4), DA was combined with one of the other pretreatment performance dimensions. Finally, we assessed a most parsimonious model, in which all indices were loaded onto one single factor (see Pretreatment Measurement Model 7 in Table 4). Each of these modified, more parsimonious measurement models yielded a significantly worse fit of the data compared to the base measurement model, as reflected in the significant  $\Delta\chi^2$  values in Table 4. We therefore concluded that DA measured a distinct dimension of pretreatment ability, and we incorporated all six dimensions of pretreatment ability into our structural model to predict word-problem outcome.

*Posttreatment Measurement Model: Are Near- and Far-Transfer Distinct Dimensions of Posttreatment Mathematics Word-Problem Skill?*

Our posttreatment measurement model included two dimensions of posttreatment word-problem skill. The first latent variable, near-transfer word problems, comprised Algorithmic Word Problems and Complex Word Problems. The second dimension, far-transfer word problems, included Real-World Word Problems, the Iowa, and WJ III Applied Problems. See Table 3 for correlations among the posttreatment measurement model latent traits. All observed variables loaded substantially and reliably onto their respective factors (standardized coefficients: .69 to .89,  $ps < .001$ ). The two-factor solution (near-transfer and far-transfer) accounted well for the data structure,  $\chi^2(4, N = 122) = 6.57, p = .154$ ; NFI = .981; NNFI = .983; CFI = .993; and SRMR = 0.0323 (see Posttreatment Measurement Model 1 in Table 4).

We contrasted this base measurement model with an alternative two-factor model comprising a simple word-problem latent variable (the Iowa and Algorithmic Word Problems) and a complex word-problem latent variable (comprised of Complex Word Problem, Real-World Word Problems, and WJ III Applied Problems). The same four independent judges who had classified near versus far transfer reclassified the measures in terms of simple versus complex. This alternative represented a theoretically different means to represent the factor structure of the word-problem outcome measures. This model fit substantially worse than the base model,  $\chi^2(4, N = 122) = 92.61, p = .0000$ ; NFI = .729; NNFI = .472; CFI = .736; and SRMR = 0.1456. (A chi-square difference score could not be used because these models are not nested.). The two-factor near- and far-transfer model included more estimated parameters than the competing measurement model; therefore, it provided the better characterization of the data and served as a basis for comparing the competing, more parsimonious model.

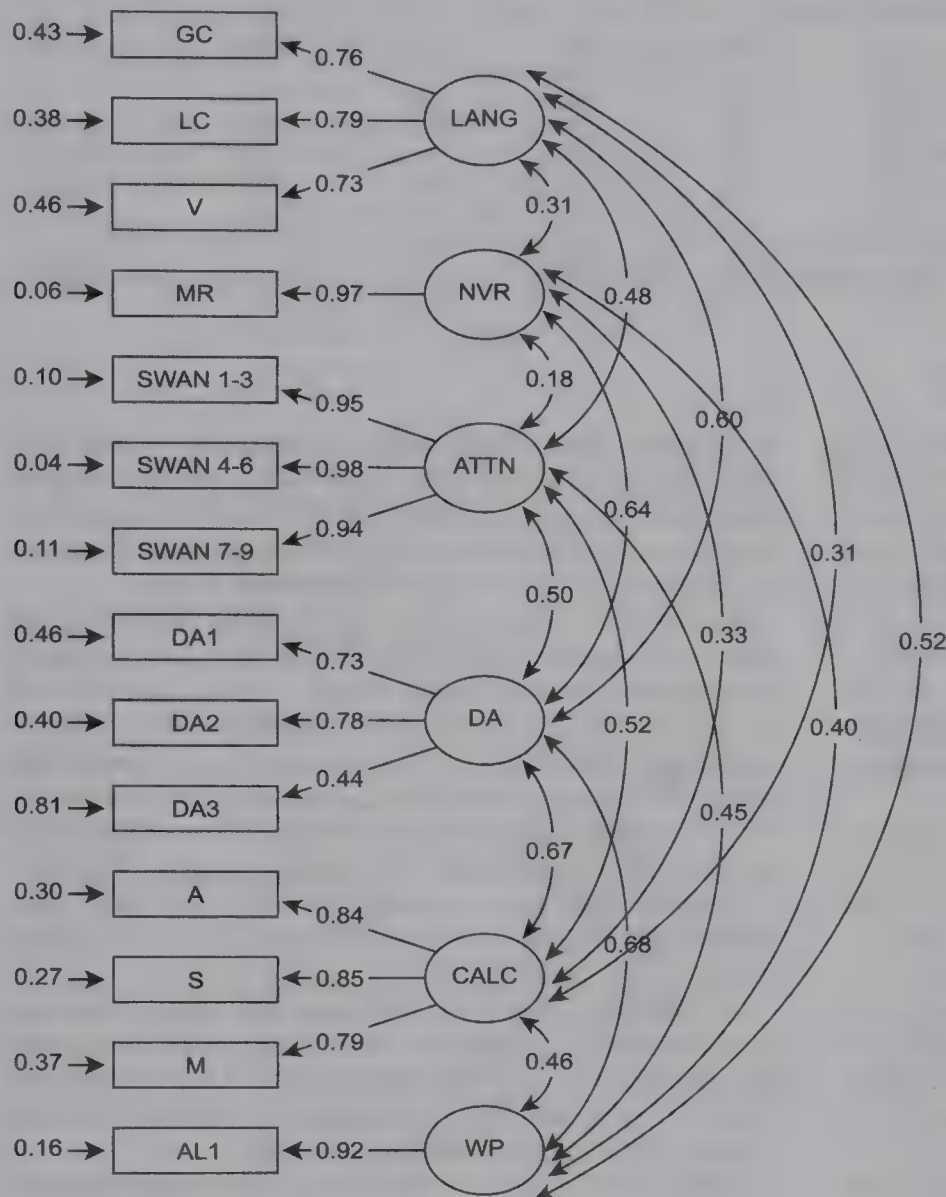
Table 4  
Fit Indices and Model Comparisons for the Competing Measurement Models

Model	df	$\chi^2$	p	NFI	NNFI	CFI	SRMR	$\Delta\chi^2$ Model 1
Pretreatment measurement models								
Six-factor model								
1. L, A, DA, C, NVR, WP	64	58.31	.6771	.965	1.000	1.000	.0389	
Five-factor models								
2. A, C, NVR, WP, L + DA	69	145.88	.0000	.927	.995	.965	.0783	87.49*
3. L, C, NVR, WP, A + DA	69	199.78	.0000	.897	.912	.933	.1325	91.47*
4. L, A, NVR, WP, C + DA	69	147.54	.0000	.926	.952	.964	.0739	89.23*
5. L, A, C, WP, NVR + DA	69	155.82	.0000	.917	.946	.954	.1803	97.51*
6. L, A, C, NVR, WP + DA	69	137.19	.0000	.931	.960	.970	.0686	78.88*
One-factor model								
7. General factor	77	634.07	.0000	.745	.731	.775	.1753	575.76*
Posttest measurement models								
Two-factor models								
1. Near transfer, far transfer	4	6.67	.1541	.981	.982	.993	.0323	
2. Simple, complex	4	92.61	.0000	.729	.472	.736	.1456	84.94*
One-factor model								
3. General math factor	5	83.41	.0000	.686	.384	.692	.1423	76.74*

Note. N = 122. L = language factor; A = attention factor; DA = dynamic assessment factor; C = calculation factor; NVR = nonverbal reasoning factor; WP = word-problem factor; NFI = norm fit index; NNFI = nonnormed fit index; CFI = comparative fit index; SRMR = standardized root-mean-square residual.

\*  $p < .001$ .

## Pretreatment Measurement Model



## Posttreatment Measurement Model

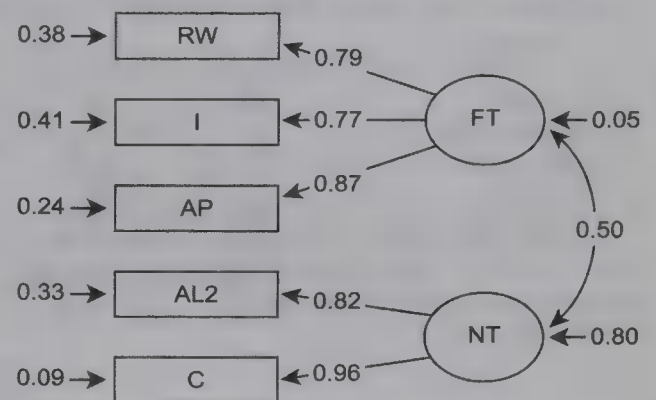


Figure 1. Pretreatment and posttest measurement models. Left: The base pretreatment measurement model, which incorporated each of the six dimensions of pretreatment ability measured in September and October: language ability (Test of Language Development—Primary Grammatical Closure [GC], Woodcock Diagnostic Reading Battery—Listening Comprehension [LC]), and Wechsler Abbreviated Scale of Intelligence [WASI] Vocabulary [V]), nonverbal reasoning (WASI Matrix Reasoning [MR]), attentive behavior (SWAN Rating Scale [SWAN 1-3, SWAN 4-6, SWAN 7-9]), Dynamic Assessment (DA1, DA2, DA3), calculations skill (Addition Fact Fluency from the Grade 3 Math Battery [A], Subtraction Fact Fluency from the Grade 3 Math Battery [S], Test of Mixed Algorithms [M]), word-problem skill (Algorithmic Word Problems [AL1]). Right: The two dimensions of posttreatment word-problem outcome measured in March: far-transfer word problems (Real-World Problem Solving [RW], the Iowa Test of Basic Skills: Problem Solving and Data Interpretation [I], and Woodcock-Johnson III Applied Problems [AP]), and near-transfer word problems (Algorithmic Word Problems [AL2] and Complex Word Problems [C]). LANG = language; NVR = nonverbal reasoning; ATTN = attention; CALC = calculation; WP = word-problem skill; FT = far-transfer word problems; NT = near-transfer word problems.

We then assessed whether both dimensions (near- vs. far-transfer) of posttreatment word-problem skill were necessary, by comparing this base measurement model against a more parsimonious, one-factor measurement model, which yielded a significantly worse fit of the data as reflected in the significant  $\Delta\chi^2$  value for Posttreatment Measurement Model 2 in Table 4. We therefore concluded that both dimensions of posttreatment word-problem skill (near- and far-transfer) were necessary, and we incorporated both dimensions into our structural model.

### Structural Model: Assessing the Predictors of Word-Problem Outcome Skill

Because few studies of this type have been conducted, we took an exploratory approach to model building. We first assessed the least parsimonious structural model. In this base model, each of the six dimensions of pretreatment performance (measured in September and October) plus treatment was included as a predictor of near-transfer word-problem skill and of far-transfer word-problem



Table 5  
*Fit Indices and Model Comparisons for the Competing Structural Models*

Model	<i>df</i>	$\chi^2$	<i>p</i>	NFI	NNFI	CFI	SRMR	$\Delta\chi^2$ Model 1
1. Full model	143	169.61	.0637	.950	.986	.989	.0758	
2. Cognitive model	149	235.10	.0000	.932	.965	.977	.0873	65.49*
3. Math model	151	210.16	.0011	.940	.976	.981	.0818	40.55*
4. DA model	153	190.91	.0203	.945	.984	.987	.0783	21.30*
5. Trimmed full model	151	174.84	.0895	.949	.998	.991	.0859	5.23

*Note.* *N* = 122. DA = dynamic assessment factor; NFI = norm fit index; NNFI = nonnormed fit index; CFI = comparative fit index; SRMR = standardized root-mean-square residual.

\* *p* < .001.

outcome (measured in March; see Table 5 and Figure 2, Full Model). In this model (and in all subsequent nested structural models), we set the correlation between treatment and the dimensions of pretreatment performance to zero, given random assignment to treatment. (The other dimensions of pretreatment performance were allowed to correlate freely.) As shown in Table 5, the model provided a good fit with the data,  $\chi^2(143, N = 122) = 169.61, p = .0637$ ; NFI = .950; NNFI = .986; CFI = .989; and SRMR = 0.0758. Treatment was the only significant predictor of near transfer. DA, pretreatment word-problem skill, and language were significant predictors of far transfer.

We then contrasted three alternative, more parsimonious structural models. In each of these models, treatment was retained as a predictor of outcome. In Models 2 and 3, we considered whether DA was in fact necessary to forecast word-problem outcomes or whether other dimensions of pretreatment ability function well without the inclusion of DA. Alternatively, it is possible that DA, in combination with treatment, is sufficient to account for the data structure in predicting word-problem outcomes, without including pretreatment cognitive abilities or pretreatment math skills. This was assessed in Model 4.

Specifically, in Model 2, the pretreatment cognitive dimensions and treatment were examined as the sole pretreatment predictors of outcome (see Cognitive Model in Table 5 and Figure 2). Given prior work establishing language (Fuchs et al., 2008; Jordan et al., 1995), nonverbal reasoning (Fuchs et al., 2005; 2006; Agness & McClone 1987), and attentive behavior (Fuchs et al., 2005; 2006; Passolunghi et al., 1999) as predictors of word-problem skill, we deemed it possible that the cognitive predictors and treatment alone would suffice in accounting for outcomes. If this Cognitive Model did not produce a significantly worse fit of the data structure compared to the base model, then assessing pretreatment foundational math skills and DA would be unnecessary, thereby providing a more parsimonious procedure for predicting outcome and undermining the importance of DA.

In Model 3, pretreatment math skills (on calculations and word problems) and treatment were considered as the sole predictors of near- and far-transfer (see Math Model in Table 5 and Figure 2). Given the transparent nature of the relation of pretreatment calculations skill and pretreatment word-problem skill to the word-problem outcomes, as well as prior work showing that pretreatment skill is often the best predictor of learning, we deemed it possible that these foundational math skills, along with treatment, would be sufficient to account for word-problem outcomes. Again,

if this Math Model did not produce a significantly worse fit of the data structure compared to the base model, then assessing the pretreatment cognitive abilities and DA would be unnecessary, consequently providing a more parsimonious approach for predicting outcome and undermining the importance of DA.

By contrast, in Model 4, we examined whether DA (along with treatment) might be sufficient to forecast near- and far-transfer word-problem outcome (see DA Model in Table 5 and Figure 2). As with the previous models, if the DA Model did not produce a significantly worse fit of the data structure compared to the base model, then assessing pretreatment cognitive abilities and pretreatment foundational skills would be unnecessary, hence providing a more parsimonious procedure for predicting outcome. This would demonstrate DA's power as a predictor of outcome even further than the base model, in which DA is one of three significant predictors of far-transfer outcome.

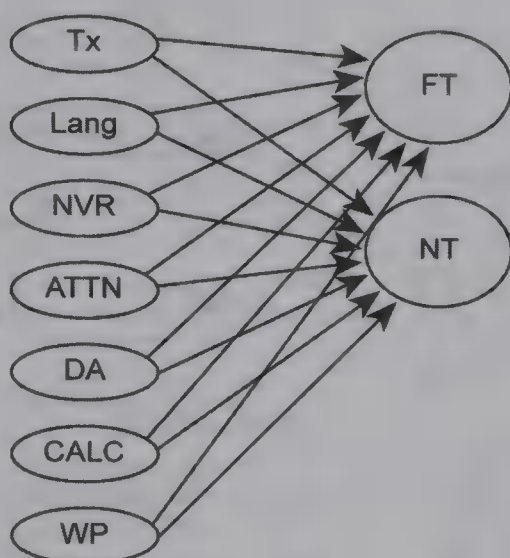
As shown in Table 5, each of these four nested, contrasting models produced a significantly worse fit of the data, suggesting that the full model is the best representation of the data structure. Figure 3 shows the full model with all coefficients specified (bolded, asterisked coefficients are *p* < .001).

Then, in the interest of reducing the parameter-to-sample-size ratio, we ran a trimmed model, deleting paths with standardized coefficients of less than .10. This trimmed model also provided a good fit with the data (see Figure 4 and last line of Table 5). Because there was no appreciable change in fit when moving from the full to the trimmed model, the simpler model is preferred. The coefficients in the full and trimmed models were highly similar, except that the path from calculations skill to near-transfer word problems increased from .14 to a significant .23 in the trimmed model.

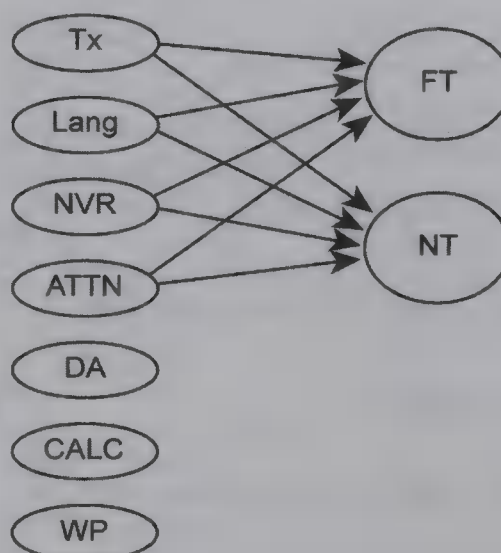
Discussion

We extended prior work on DA in several ways. First, earlier studies formulated DA tasks to address general cognitive abilities (e.g., Day et al., 1997; Speece et al., 1990) or a cognitive ability specifically linked to performance in the academic domain (e.g., Spector, 1992; Swanson & Howard, 2005). In the present study, by contrast, we centered DA on students' potential to learn algebra, which is neither a general cognitive ability nor a precursor to third-grade word-problem skill. In fact, skill with algebra is not even considered foundational to third-grade word-problem skill and instead represents novel academic content within the same broad academic area (mathematics). The potential advantage of

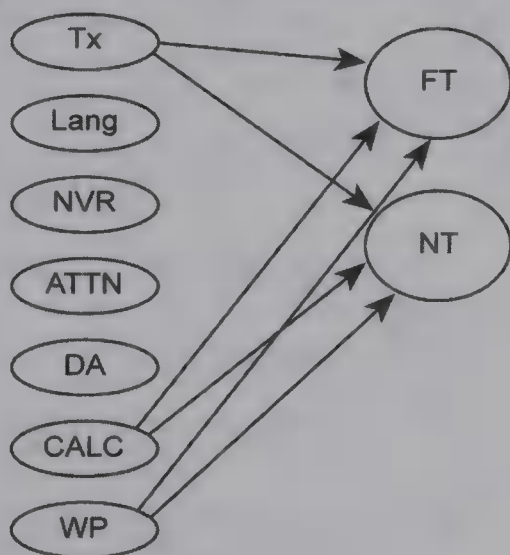
## 1: Full Model



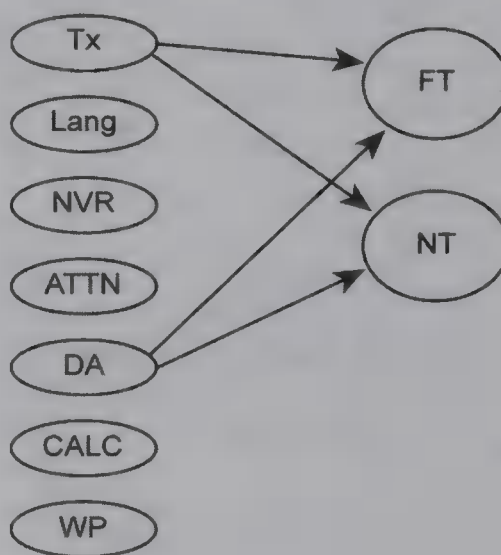
## 2: Cognitive Model



## 3: Math Model



## 4: DA Model



*Figure 2.* Four structural models predicting pretreatment performance (measured in September and October) plus treatment to posttreatment near- and far-transfer outcomes (measured in March). The Full (base) Model is the least parsimonious structural, in which each of the six dimensions of pretreatment performance plus treatment was included as a predictor of outcomes. Each of the subsequent (nested) models is more parsimonious. In the Cognitive Model, pretreatment cognitive dimensions and treatment were examined as the sole predictors of outcomes. In the Math Model, pretreatment calculations and word-problem skill and treatment were considered as the sole predictors of outcomes. In the dynamic assessment (DA) Model, pretreatment DA and treatment were examined as the sole predictors of outcomes. Tx = treatment; Lang = language (Test of Language Development—Primary Grammatic Closure, Woodcock Diagnostic Reading Battery—Listening Comprehension, and Wechsler Abbreviated Scale of Intelligence [WASI] Vocabulary); NVR = nonverbal reasoning (WASI Matrix Reasoning); ATTN = attention (SWAN Rating Scale); DA = dynamic assessment; CALC = calculations skill (Addition Fact Fluency from the Grade 3 Math Battery, Subtraction Fact Fluency from the Grade 3 Math Battery, Test of Mixed Algorithms); WP = word-problem skill (Algorithmic Word Problems); FT = far-transfer word problems (Real-World Problem Solving, the Iowa Test of Basic Skills: Problem Solving and Data Interpretation, and Woodcock-Johnson III Applied Problems); NT = near-transfer word problems (Algorithmic Word Problems and Complex Word Problems).

using a DA task with novel academic content, such as algebra at third grade, is that it increases the probability that students' DA performance is attributable to their potential to learn new content rather than to existing ability to perform the DA task. In fact,

among the 122 participants, only 2 students exhibited mastery on all three DA skills prior to the introduction of scaffolded instruction (another 2 exhibited mastery on two of the three skills prior to scaffolded instruction, and another 34 exhibited mastery on one of



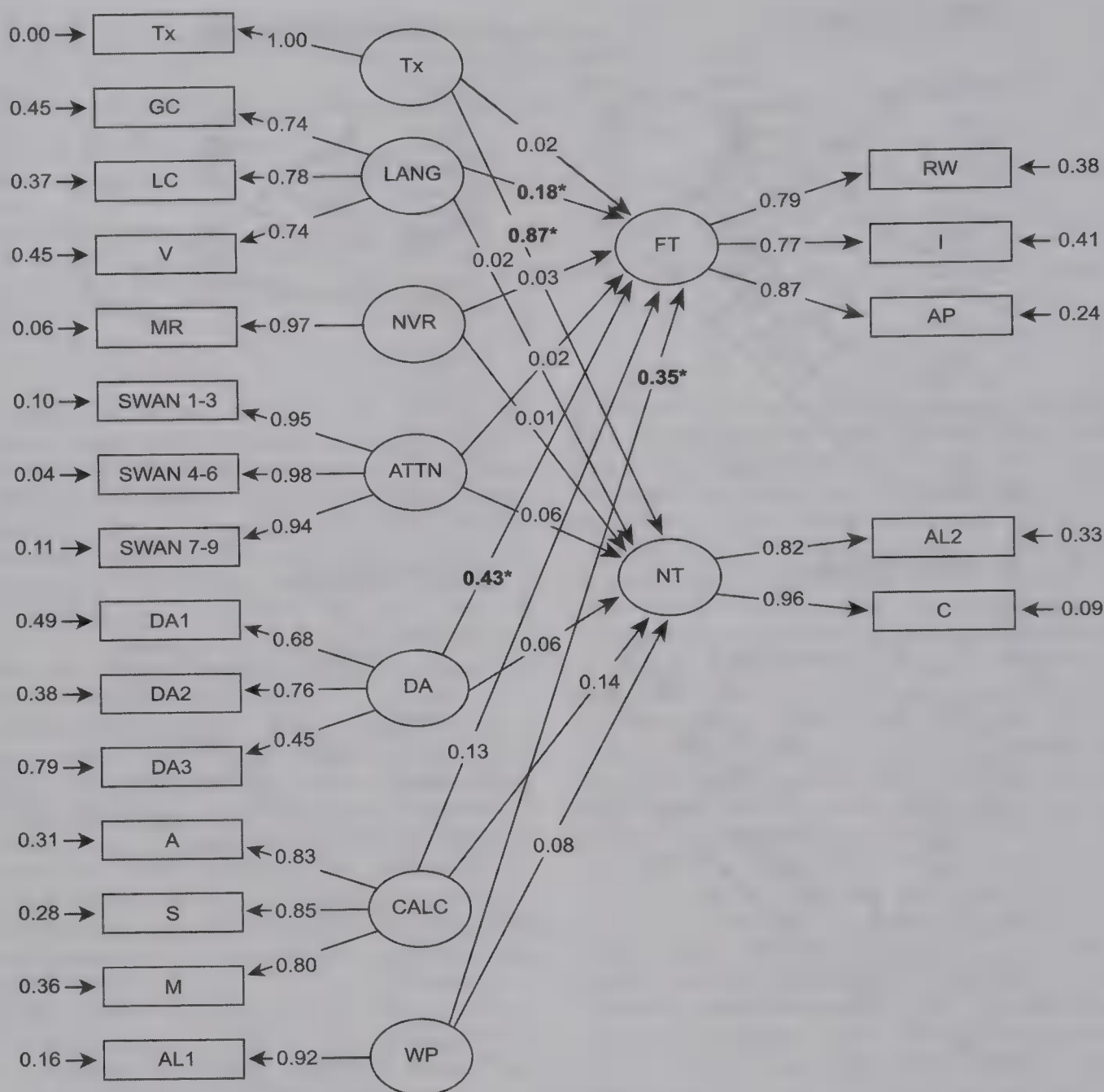


Figure 3. The full SEM with all coefficients specified. Left: The six dimensions of pretreatment ability measured in September and October are language ability (Test of Language Development—Primary Grammatical Closure [GC], Woodcock Diagnostic Reading Battery—Listening Comprehension [LC]), and Wechsler Abbreviated Scale of Intelligence [WASI] Vocabulary [V]), nonverbal reasoning (WASI Matrix Reasoning [MR]), attentive behavior (SWAN Rating Scale [SWAN 1-3, SWAN 4-6, SWAN 7-9]), Dynamic Assessment (DA1, DA2, DA3), calculations skill (Addition Fact Fluency from the Grade 3 Math Battery [A], Subtraction Fact Fluency from the Grade 3 Math Battery [S], Test of Mixed Algorithms [M]), word-problem skill (Algorithmic Word Problems [AL1]). Middle: Tx = treatment; Lang = language (Test of Language Development—Primary Grammatical Closure, Woodcock Diagnostic Reading Battery—Listening Comprehension, and WASI Vocabulary); NVR = nonverbal reasoning (WASI Matrix Reasoning); ATTN = attention (SWAN Rating Scale); DA = dynamic assessment; CALC = calculations skill (Addition Fact Fluency from the Grade 3 Math Battery, Subtraction Fact Fluency from the Grade 3 Math Battery, Test of Mixed Algorithms); WP = word-problem skill (Algorithmic Word Problems); FT = far-transfer word problems (Real-World Problem Solving, the Iowa Test of Basic Skills: Problem Solving and Data Interpretation, and Woodcock-Johnson III Applied Problems); NT = near-transfer word problems (Algorithmic Word Problems and Complex Word Problems). Right: The two dimensions of posttreatment word-problem outcome measured in March: far-transfer word problems (Real-World Problem Solving [RW], the Iowa Test of Basic Skills: Problem Solving and Data Interpretation [I], and Woodcock-Johnson III Applied Problems [AP]), and near-transfer word problems (Algorithmic Word Problems [AL2] and Complex Word Problems [C]). \*  $p < .001$ .

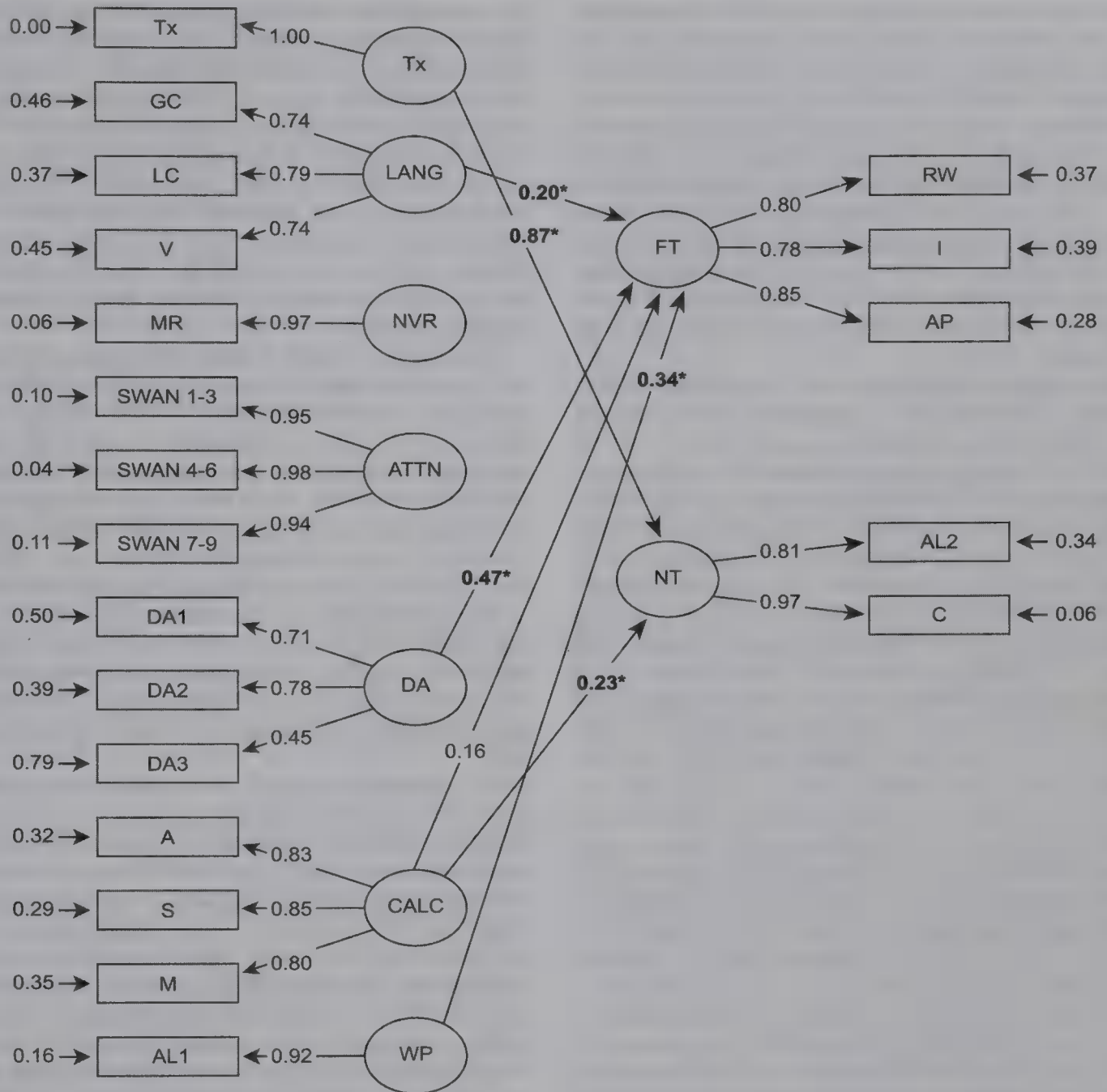


Figure 4. The trimmed SEM. Left: The six dimensions of pretreatment ability measured in September and October are language ability (Test of Language Development—Primary Grammatical Closure [GC], Woodcock Diagnostic Reading Battery—Listening Comprehension [LC]), and Wechsler Abbreviated Scale of Intelligence [WASI] Vocabulary [V]), nonverbal reasoning (WASI Matrix Reasoning [MR]), attentive behavior (SWAN Rating Scale [SWAN 1-3, SWAN 4-6, SWAN 7-9]), Dynamic Assessment (DA1, DA2, DA3), calculations skill (Addition Fact Fluency from the Grade 3 Math Battery [A], Subtraction Fact Fluency from the Grade 3 Math Battery [S], Test of Mixed Algorithms [M]), word-problem skill (Algorithmic Word Problems [AL1]). Middle: Tx = treatment; Lang = language (Test of Language Development—Primary Grammatical Closure, Woodcock Diagnostic Reading Battery—Listening Comprehension, and WASI Vocabulary); NVR = nonverbal reasoning (WASI Matrix Reasoning); ATTN = attention (SWAN Rating Scale); DA = dynamic assessment; CALC = calculations skill (Addition Fact Fluency from the Grade 3 Math Battery, Subtraction Fact Fluency from the Grade 3 Math Battery, Test of Mixed Algorithms); WP = word-problem skill (Algorithmic Word Problems); FT = far-transfer word problems (Real-World Problem Solving, the Iowa Test of Basic Skills: Problem Solving and Data Interpretation, and Woodcock-Johnson III Applied Problems); NT = near-transfer word problems (Algorithmic Word Problems and Complex Word Problems). Right: The two dimensions of posttreatment word-problem outcome measured in March: far-transfer word problems (Real-World Problem Solving [RW], the Iowa Test of Basic Skills: Problem Solving and Data Interpretation [I], and Woodcock-Johnson III Applied Problems [AP]), and near-transfer word problems (Algorithmic Word Problems [AL2] and Complex Word Problems [C]). \*  $p < .001$ .



the three skills prior to scaffolded instruction). The remaining 84 students failed to exhibit mastery on any of the skills prior to scaffolded instruction. At the same time, it is important that the novel academic content connects in some important ways to the predicted learning outcome. In designing this study's DA as a method of forecasting development of word-problem skill, we hypothesized a connection between algebraic cognition and word-problem skill because (a) although linguistic content is absent from algebra, algebra does require understanding of relations among quantities, as is the case for solving word problems, and (b) some research (e.g., Fuchs, Seethaler, et al., 2008) shows that algebra can be used to develop students' word-problem skill, even in the primary grades.

In addition, we extended prior DA work in terms of the scope of the variables we considered as competitors to DA in predicting outcome. That is, we incorporated the nature of treatment into the prediction model, even as we simultaneously included static cognitive abilities as well as foundational math skill (on calculations as well as word problems) to predict outcomes that were proximal to as well as distal from the instruction students received in school. Our pretreatment measurement modeling supported DA as representing a dimension of pretreatment ability distinct from existing language ability, nonverbal reasoning, attentive behavior, and math skill. This finding echoes the work of Day et al. (1997) conducted with preschoolers, who found that the measurement model that retained DA as a construct separate from pre-DA provided better fit of the data. Together, Day et al.'s and our pretreatment measurement modeling indicates that DA may represent an aspect of students' learning potential that is distinct from conventional static measures. This makes sense because static assessment quantifies already-developed abilities that are influenced by environmental factors such as educational opportunity in school, parental support, and test-taking skills (Grigorenko & Sternberg, 1998), whereas DA measures a student's potential to learn. In this way, findings provide support for Vygotsky's (1978) concept of the zone of proximal development, or the distance between a child's realized developmental level as assessed by independent performance and that child's potential development as assessed via supported performance, as a distinct cognitive characteristic.

Importantly, however, our work also extends Day et al.'s (1997) investigation as well as other seminal DA studies that examine DA's utility as a predictor of academic performance (e.g., Speece et al., 1990; Swanson & Howard, 2005) by assessing whether DA is a viable predictor of future (rather than concurrent) achievement. It also extends Spector's (1992) prediction of kindergarteners' future reading performance by considering how instruction affects relations among static and dynamic measures in forecasting school learning. In terms of whether DA is a viable predictor of future learning, our results provide support. We contrasted a full model of predictors, which included treatment plus six dimensions of pretreatment performance (language ability, nonverbal reasoning, attentive behavior, DA, calculations skill, and word-problem skill), against models that isolated static cognitive abilities (language ability, nonverbal reasoning, and attentive behavior), that isolated DA, and that isolated math skill (calculations and word problems). None of these contrasting, more parsimonious models fit the data as well as the full model. Therefore, although DA (along with treatment) was insufficient, it was necessary in accounting for the

data structure. This finding suggests that DA may serve an important role in predicting students' future learning. It also indicates that forecasting later academic performance creates variance for DA to capture. This may occur because upcoming learning may relate better with learning potential as indexed via DA than with static assessments that are determined in part by culture, socioeconomic, and previous learning opportunity. More specifically, these findings support a possible connection between algebraic cognition and word-problem skill, a possibility that should be pursued in future work, and indicate that novel academic content may provide a productive avenue for designing DAs in other academic domains.

With respect to how instruction affects relations among static and dynamic measures in forecasting actual school learning, our measurement models showed that two latent variables of outcome, near- and far-transfer, were necessary to account for the structure of students' posttreatment word-problem skill, and both dimensions of outcome were important for understanding how instruction affects the relations among static and dynamic measures in forecasting learning. Although all items on all five outcome measures were novel and none had been used for instruction, only treatment and pretreatment calculations skill accounted for learning on the latent outcome variable more closely aligned with treatment. By contrast, to forecast learning on word-problem measures more distally related to treatment, DA as well as pretreatment language ability and pretreatment word-problem skill were necessary.

That treatment forecasted near- but not far-transfer word-problem outcome is not surprising because randomized control trials (e.g., Fuchs et al., in press; Fuchs, Hamlett, & Powell, 2003) have, while demonstrating the efficacy of the validated word-problem treatment used in the present study, revealed stronger effect sizes for near-transfer word-problem outcome measures (e.g., 1.73–1.90; Fuchs et al., in press) than for far-transfer word-problem outcome measures (e.g., 0.36; Fuchs et al., in press). It is possible that with more comprehensive treatments that more fully address more generalized performance, the value of DA (as well as language ability and pretreatment word-problem skill) in predicting outcome would decrease, as the value of the treatment increases. However, effecting transfer to distal problems represents a formidable challenge in the area of math problem solving (Bransford & Schwartz, 1999; Cooper & Sweller, 1987; Mayer, Quilici, & Moreno, 1999). Moreover, we note that the validated intervention employed in the present study is theoretically rooted in schema theory that addresses transfer (e.g., Chi, Feltovich, & Glaser, 1981; Gick & Holyoake, 1980; Mayer, 1992; Quilici & Mayer, 1996) and has been shown to accomplish transfer more effectively than conventional math problem-solving instruction (Fuchs et al., 2003a). In addition, it is impossible to address all novel problem types within any given instructional program. Thus, the prediction of distal transfer, as reflected in the kinds of complex, real-world problem solving and high-stakes assessments used as distal measures in the present study, is important.

Results are relevant to the present education reform known as responsiveness-to-instruction (RTI). RTI is embedded in a multi-level prevention system, borrowed from the public health system. In a multilevel prevention system, general education constitutes primary prevention. Students who do not respond to this universal, core program enter secondary prevention. In most research, this



involves one or more rounds of small-group tutoring using a validated tutoring protocol. Students who do not respond to this more intensive secondary prevention program are understood to have demonstrated “unexpected failure” to validated instruction (to which most students respond). On this basis, unresponsive students are deemed appropriate for more intensive and individualized forms of instruction at the tertiary prevention level. In this way, RTI has two complementary purposes: (a) to identify at-risk students early and offer them intervention to prevent the onset of severe and often intractable deficits and (b) to identify students who are unresponsive to standard, validated instruction and who therefore require individualized instruction. In most RTI systems, these chronically unresponsive students are considered to have a learning disability (LD). In fact, the 2004 reauthorization of the Individuals with Disabilities Education Improvement Act of 2004 encourages RTI for LD identification as an alternative to the traditional identification procedure, which requires documentation of a discrepancy between IQ and school achievement and often delays identification until the intermediate grades (Vaughn & Fuchs, 2003).

Within RTI models, responsiveness to secondary prevention is considered the “test” for differentiating between two explanations for low achievement: inadequate instruction versus disability. If a child responds poorly to instruction that benefits most students, then this eliminates instructional quality as an explanation of poor learning and instead suggests disability. Although earlier identification/treatment of LD represents a potential advantage of RTI, conducting secondary preventive tutoring requires at least 10 weeks and, in some models, as many as 30 weeks. Present findings suggest that DA might serve an important function in predicting responsiveness to intervention, by forecasting distal, critical aspects of learning that are not highly relevant to the student’s proximal instructional needs. For example, within RTI reading models at first grade, secondary prevention focuses predominantly on the development of word-level skills and fluency because some work (e.g., Fuchs & Fuchs, 2005; Fuchs, Fuchs, Kazdan, & Allen, 1999) suggests that instruction to develop comprehension strategies when students are still struggling to identify words accurately and fluently may detract from reading development. Nonetheless, studies (Catts & Hogan, 2002; Compton, Fuchs, Fuchs, Elleman, & Gilbert, in press; Leach, Scarborough, & Rescorla, 2003; Lipka, Lesaux, & Siegel, 2006) show that a small but significant portion of the population experiences late-emerging reading disability, whereby their word-level and fluency skills develop typically, but serious comprehension deficits become evident at third or fourth grade. Yet, research has not reliably identified early predictors of late-emerging reading disability (Compton et al., in press). It is possible that DA might be used in first grade to forecast development of late-emerging reading disability. This would help teachers design appropriate instruction for this subset of students early on. In a different way, DA might be used productively within an RTI framework to help identify students who will ultimately, 10–30 weeks later, prove unresponsive to secondary prevention. If earlier identification, via DA, were possible, then more intensive and individualized forms of instruction could be introduced sooner, without children experiencing 10–30 weeks of additional failure.

Before closing, we note the small parameter-to-sample-size ratio in this study. Use of maximum likelihood estimation with small samples size has a tendency to produce chi-square estimates that

are too large. Thus, the decision for accepting or rejecting a particular model may vary as a function of sample size. In addition, sample size affects power in a predictable way, with smaller samples sizes having less power to detect differences between competing models using the chi-square statistic. Although our sample size was smaller than desirable, we did succeed in detecting significant differences using the  $\Delta\chi^2$  statistic across competing models. In addition, our reported Type 2 and Type 3 fit indices reflected adequate fit for our final measurement and structural models. These indices have been shown to be substantially less biased by sample size (Hu & Bentler, 1998). Therefore, we have some confidence that even with our sample size, the final models represent a good estimation of the relations among measures. Moreover, we ran a trimmed model, deleting paths with standardized coefficients of less than .10, which provided a similarly good fit with the data.

Nevertheless, the parameter-to-sample-size ratio remains a limitation of the present study, and future DA research employing structural equation modeling should incorporate larger samples. With this caution in mind, we conclude that findings provide an important basis for additional research assessing the utility of DA for indexing distal outcomes and for identifying students who require tertiary prevention in a more timely way. The present database indicates the potential utility of DA for predicting math problem-solving generalization across third grade. Parallel work, with larger samples, appears warranted to examine other aspects of academic learning.

## References

- Agness, P. J., & McClone, D. G. (1987). Learning disabilities: A specific look at children with spina bifida. *Insights*, 9–9.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Bransford, J. D., Delclos, V. R., Vye, N. J., Burns, M. S., & Hasselbring, T. S. (1987). State of the art and future directions. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 479–496). New York: Guilford Press.
- Bransford, J. D., & Schwartz, D. L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P. D. Pearson (Eds.), *Review of research in education* (pp. 61–100). Washington, DC: American Educational Research Association.
- Budoff, M. (1967). Learning potential among institutionalized young adult retardates. *American Journal of Mental Deficiency*, 72, 404–411.
- Byrne, B., Fielding-Barnsley, R., & Ashley, L. (2000). Effects of preschool phoneme identity training after six years: Outcome level distinguished from rate of response. *Journal of Educational Psychology*, 92, 659–667.
- Campione, J. C. (1989). Assisted testing: A taxonomy of approaches and an outline of strengths and weaknesses. *Journal of Learning Disabilities*, 22, 151–165.
- Campione, J. C., & Brown, A. L. (1987). Linking dynamic assessment with school achievement. In C. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 82–115). New York: Guilford Press.
- Catts, H. W., & Hogan, T. P. (2002, June). *The fourth grade slump: Late emerging poor readers*. Poster presented at the annual conference of the Society for the Scientific Study of Reading, Chicago, IL.
- Chard, D. J., Clarke, B., Baker, S., Otterstedt, J., Braun, D., & Katz, R. (2005). Using measures of number sense to screen for difficulties in mathematics: Preliminary findings. *Assessment for Effective Intervention*, 30(2), 3–14.
- Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and



- representation of physics problems by experts and novices. *Cognitive Science*, 5, 121–152.
- Clarke, B., & Shinn, M. R. (2004). A preliminary investigation into the identification and development of early mathematics curriculum-based measurement. *School Psychology Review*, 33, 234–248.
- Compton, D. L., Fuchs, D., Fuchs, L. S., Elleman, A. M., & Gilbert, J. K. (in press). Tracking children who fly below the radar: Latent transition modeling of students with late-emerging reading disability. *Learning and Individual Differences*.
- Cooper, G., & Sweller, J. (1987). Effects of schema acquisition and rule automation on mathematical problem solving transfer. *Journal of Educational Psychology*, 79, 347–362.
- CTB/McGraw-Hill. (2003). *TerraNova technical manual*. Monterey, CA: Author.
- Day, J. D., Engelhardt, J. L., Maxwell, S. E., & Bolig, E. E. (1997). Comparison of static and dynamic assessment procedures and their relation to independent performance. *Journal of Educational Psychology*, 89, 358–368.
- Ferrara, R. A., Brown, A. L., & Campione, J. C. (1986). Children's learning and transfer of inductive reasoning rules: Studies of proximal development. *Child Development*, 57, 1087–1099.
- Feuerstein, R. (1979). *The dynamic assessment of retarded performers. The Learning Potential Assessment Device, theory, instruments, and techniques*. Baltimore: University Park Press.
- Fuchs, D., & Fuchs, L. S. (2005). Peer-Assisted Learning Strategies: Promoting word recognition, fluency, and comprehension in young children. *Journal of Special Education*, 39, 34–44.
- Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology*, 97, 493–513.
- Fuchs, L. S., Fuchs, D., Compton, D. L., Powell, S. R., Seethaler, P. M., Capizzi, A. M., et al. (2006). The cognitive correlates of third-grade skill in arithmetic, algorithmic computation, and arithmetic word problems. *Journal of Educational Psychology*, 98, 29–43.
- Fuchs, L. S., Fuchs, D., Craddock, C., Hollenbeck, K. N., Hamlett, C. L., & Schatschneider, C. (in press). Effects of small-group tutoring with and without validated classroom instruction on at-risk students' math problem solving: Are two tiers of prevention better than one? *Journal of Educational Psychology*.
- Fuchs, L. S., Fuchs, D., Finelli, R., Courey, S. J., & Hamlett, C. L. (2004). Expanding schema-based transfer instruction to help third graders solve real-life mathematical problems. *American Educational Research Journal*, 41, 419–445.
- Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C. L., Dutka, S., & Katzaroff, M. (2000). The importance of providing background information on the structure and scoring of performance assessments. *Applied Measurement in Education*, 13, 1–34.
- Fuchs, L. S., Fuchs, D., Kazdan, S., & Allen, S. (1999). Effects of peer-assisted learning strategies in reading with and without training in elaborated help giving. *Elementary School Journal*, 99, 201–220.
- Fuchs, L. S., Fuchs, D., Prentice, K., Burch, M., Hamlett, C. L., Owen, R., Hosp, M., & Janeck, D. (2003a). Explicitly teaching for transfer: Effects on third-grade students' mathematical problem solving. *Journal of Educational Psychology*, 95, 293–304.
- Fuchs, L. S., Fuchs, D., Prentice, K., Burch, M., Hamlett, C. L., Owen, R., & Schroeter, K. (2003b). Enhancing third-grade students' mathematical problem solving with self-regulated learning strategies. *Journal of Educational Psychology*, 95, 306–315.
- Fuchs, L. S., Fuchs, D., Prentice, K., Hamlett, C. L., Finelli, R., & Courey, S. J. (2004). Enhancing mathematical problem solving among third-grade students with schema-based instruction. *Journal of Educational Psychology*, 96, 635–647.
- Fuchs, L. S., Fuchs, D., Stuebing, K., Fletcher, J. M., Hamlett, C. L., & Lambert, W. E. (2008). Problem-solving and computational skills: Are they shared or distinct aspects of mathematical cognition? *Journal of Educational Psychology*, 100, 30–47.
- Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1990). *Test of Computational Fluency*. Available from L. S. Fuchs, 328 Peabody, Vanderbilt University, Nashville, TN 37203.
- Fuchs, L. S., Hamlett, C. L., & Powell, S. R. (2003). *Grade 3 Math Battery*. Available from L. S. Fuchs, 328 Peabody, Vanderbilt University, Nashville, TN 37203.
- Fuchs, L. S., Seethaler, P. M., Powell, S. R., Fuchs, D., Hamlett, C. L., & Fletcher, J. M. (2008). Effects of preventative tutoring on the mathematical problem solving of third-grade students with math and reading difficulties. *Exceptional Children*, 74, 155–173.
- Gick, M. L., & Holyoake, K. J. (1980). Analogical problem solving. *Cognitive Psychologist*, 12, 306–355.
- Greenes, C., Larson, M., Leiva, M. A., Shaw, J. M., Stiff, L., Vogeli, B. R., & Yeatts, K. (2007). *Houghton Mifflin math*. Boston: Houghton Mifflin.
- Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin*, 124, 85–111.
- Hoover, H. D., Dunbar, S. B., Frisbie, D. A. (2001). *Iowa Test of Basic Skills*. Rolling Meadows, IL: Riverside.
- Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparametrized model specification. *Psychological Methods*, 3, 424–453.
- Individuals with Disabilities Education Improvement Act of 2004, Pub. L. No. 108–446, 118 Stat. 2647 (2004)
- Jaccard, J., & Wan, C. K. (1996). *LISREL approaches to interaction effects in multiple regression*. Thousand Oaks, CA: Sage.
- Jordan, N. C., Levine, S. C., & Huttenlocher, J. (1995). Calculation abilities in young children with different patterns of cognitive functioning. *Journal of Learning Disabilities*, 28, 53–64.
- Jöreskog, K. G., & Sörbom, D. (2004). *LISREL 8.7 for Windows* [Computer software]. Lincolnwood, IL: Scientific Software International.
- Kansas State Board of Education. (1991). *Kansas Quality Performance Accreditation*. Topeka, KS: Author.
- Kern, B. (1930). *Wirkungsformen der Übung* [Effects in training]. Munster, Germany: Helios.
- Kirk, S., McCarthy, J., & Kirk, W. (1968). *Examiner's manual: Illinois Test of Psycholinguistic Ability Grammatical Closure*. Urbana: Illinois University Press.
- Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford Press.
- Leach, J. M., Scarborough, H. S., & Rescorla, L. (2003). Late-emerging reading disabilities. *Journal of Educational Psychology*, 95, 211–224.
- Lembke, E., & Foegen, A. (2006, February). *Monitoring student progress in early math*. Paper presented at the 14th annual meeting of the Pacific Coast Research Conference, San Diego, CA.
- Lipka, O., Lesaux, N. K., & Siegel, L. S. (2006). Retrospective analyses of the reading development of Grade 4 students with reading disabilities: Risk status and profiles over 5 years. *Journal of Learning Disabilities*, 39, 364–378.
- MacCallum, R. C. (1995). Model specification: Procedures, strategies, and related issues. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 16–36). Thousand Oaks, CA: Sage.
- Mayer, R. E. (1992). *Thinking, problem solving, cognition* (2nd ed.). New York: Freeman.
- Mayer, R. E., Quilici, J. L., & Moreno, R. (1999). What is learned in an after-school computer club? *Journal of Educational Computing Research*, 20, 223–235.
- McGrew, K., & Woodcock, R. W. (2001). *Woodcock-Johnson Psycho-Educational Battery—III: Technical manual*. Chicago: Riverside.
- Murray, B. A., Smith, K. A., & Murray, G. G. (2000). The test of phoneme

- identities: Predicting alphabetic insight in pre-alphabetic readers. *Journal of Literacy Research*, 32, 421–477.
- Newcomer, P. L., & Hammill, D. D. (1988). *Test of Language Development* (Rev. Ed.). Austin, TX: Pro-Ed.
- Passolunghi, M. C., Cornoldi, C., & De Liberto, S. (1999). Working memory and inhibition of irrelevant information in poor problem solvers. *Memory & Cognition*, 27, 779–790.
- Penrose, L. S. (1934). *Mental defect*. New York: Farrar and Rinehart.
- Psychological Corporation. (1999). *Wechsler Abbreviated Scale of Intelligence*. San Antonio, TX: Harcourt Brace & Company.
- Quilici, J. L., & Mayer, R. E. (1996). Role of examples in how students learn to categorize statistics word problems. *Journal of Educational Psychology*, 88, 144–161.
- Raven, J. C. (1960). *Standard progressive matrices, sets A, B, C, D, and E*. Cambridge, England: Lewis and Co.
- Rey, A. (1934). D'un procede pour evaluer l'educabilite [A method for assessing educability]. *Archive de Psychologie*, 24, 297–337.
- Rivera-Batiz, F. L. (1992). Quantitative literacy and the likelihood of employment among young adults in the United States. *Journal of Human Resources*, 27, 313–328.
- Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.
- Spector, J. E. (1992). Predicting progress in beginning reading: Dynamic assessment of phonemic awareness. *Journal of Educational Psychology*, 84, 353–363.
- Speece, D. L., Cooper, D. H., & Kibler, J. M. (1990). Dynamic assessment, individual differences, and academic achievement. *Learning and Individual Differences*, 2, 113–127.
- Sternberg, R. J. (1996). *Successful intelligence*. New York: Simon & Schuster.
- Swanson, H. L., & Howard, C. B. (2005). Children with reading disabilities: Does dynamic assessment help in the classification? *Learning Disability Quarterly*, 28, 17–34.
- Swanson, J., Schuck, S., Mann, M., Carlson, C., Hartman, K., Sergeant, J., et al. (2004). Categorical and dimensional definitions and evaluations of symptoms of ADHD: The SNAP and the SWAN rating scales. Retrieved December 20, 2004, from [http://www.adhd.net/SNAP\\_SWAN.pdf](http://www.adhd.net/SNAP_SWAN.pdf)
- Tzuriel, D., & Haywood, H. C. (1992). The development of interactive-dynamic approaches for assessment of learning potential. In H. C. Haywood & D. Tzuriel (Eds.), *Interactive assessment* (pp. 3–37). New York: Springer-Verlag.
- Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to intervention: The promise and potential problems. *Learning Disabilities Research and Practice*, 18, 137–146.
- Vygotsky, L. S. (1962). *Thought and language*. Cambridge, MA: MIT Press (Original work published 1934).
- Vygotsky, L. S. (1978). *Mind and society*. Cambridge, MA: Harvard University Press.
- Wechsler, D. (1991). *Wechsler Intelligence Scale for Children* (3rd rev.). San Antonio, TX: Harcourt.
- West, S. G., Finch, J. F., & Curran, P. J. (1994). Structural equation models with nonnormal variables. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56–75). Thousand Oaks, CA: Sage.
- Wilkinson, G. S. (1993). *Wide Range Achievement Test 3*. Wilmington, DE: Wide Range.
- Woodcock, R. W. (1997). *Woodcock Diagnostic Reading Battery*. Itasca, IL: Riverside.
- Woodcock, R. W. (1998). *Woodcock Reading Mastery Tests—Revised*. Circle Pines, MN: American Guidance Service.
- Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-Educational Battery—Revised*. Allen, TX: DLM Teaching Resources.
- Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Cognitive Abilities*. Itasca, IL: Riverside.
- Zhu, J. (1999). *WASI manual*. San Antonio, TX: Psychological Corporation.

## Appendix

### Instructional Scaffolding Levels for DA Skill A

For DA Skill A, during *Instructional Scaffolding Level 1*, the tester works problems for the student without explanation. The tester says, “I’m going to work some problems for you. Watch carefully while I work because I’ll ask you to do some more problems like these in a few minutes.” The tester then demonstrates three problems. Then the tester says, “Now you try to do some problems like these” and readministers the six-item test.

During *Instructional Scaffolding Level 2*, the tester explains the plus and equal signs mean and what the problem means/is asking. The tester says, “The plus sign tells us to add (*point*)”. And the problem asks us to figure out what number *x* stands for. What number (*point*) can we add to 6 (*point*) to make the total 8 (*point*)? *If correct*: “That’s right! 2 + 6 adds to 8. So the missing

number is 2. I write *x* equals 2 (*write*). Now let’s look at another problem.” *If incorrect*: “That’s not quite right. 2 plus 6 equals 8. So the missing number is 2. I write *x* equals 2 (*write*). Let’s try another problem. Now try to do this practice problem (*guide student using same procedure*).”

During *Instructional Scaffolding Level 3*, the tester elaborates on what *x* means, relating *x* to the “blank” that is used more commonly in the primary grades. The tester explains, “The blank (*point*) stands for the missing number, just like the *x* stands for the missing number in this problem (*point*). These two problems mean the same thing. They ask us to figure out what number *x*, or the blank, stands for. The problem asks, “What number (*point*) can we add to 6 (*point*) to make the total 14 (*point*)?” 6 plus what number equals 14?” *If correct*: “That’s right! 8 + 6 equals 14. So I write



8 in the blank. That tells me that the missing number,  $x$ , equals 8.”  
*If incorrect:* “That’s not quite right. 6 (*point*) plus 8 equals 14. So, the missing number is 8. So I write 8 in the blank. That tells me that the missing number,  $x$ , equals 8. After you figure out what the missing number is, you write the missing number like this:  $x = 8$ . We read the problem like this:  $6 + x = 14$ ;  $x = 8$ . Let’s try another problem, but this time you’ll try to do the work” (*guide student using same procedures*).

During *Instructional Scaffolding Level 4*, the tester provides the student with the min counting strategy for finding the missing number (i.e., starting with the larger addend and then counting up the second addend to derive the sum). The tester says, “What number (*point*) do we add to  $x$  (*point*)?” (Student: 7) “Yes, this problem tells us, 7 plus  $x$ . Put the number being added to  $x$ , 7, in your head. Now count up to the total, 13 (*point*). When I count up, I use my fingers until I reach the total (*point*), like this. I already have 7 in my head (*point to head*). I start counting up using the *next* number: 8 (*put up one finger*), 9 (*put up second finger, and continue putting up fingers for each subsequent number*), 10, 11, 12, 13. How many fingers am I holding up?” (Student: 6). “Yes, I have 6 fingers up. So,  $x$  equals 6. I know that 7 (*point*) + 6 (*write 6 above the x*) equals 13 (*point*). So, I know that  $x$  equals 6, and I write my answer like this:  $x = 6$ . You write the answer for me (*student writes*).” *If correct:* “That’s right! # + # equals #. So I write # in the blank. That tells me that the missing number,  $x$ , equals #.” *If incorrect:* “That’s not quite right. # (*point*) plus # equals #. So, the missing number is #. So I write # in the blank. That tells me that the missing number,  $x$ , equals #. Let’s try another one, but this time you’ll try to do it on your own” (*guide student using same procedures*).

During *Instructional Scaffolding Level 5*, the tester adds color coding to help the student discriminate the salient components of

the min strategy, thereby increasing the level of scaffolding with concreteness. The tester says, “This problem asks us to find what number  $x$  stands for. The problem asks, ‘What number (*point*) can we add to 3 (*point*) to make the total 5 (*point*)?’ What number (*point*) do we add to  $x$  (*point*)?” (Student: 3) “Yes, this problem tells us, 3 plus  $x$ . This number is written in green to help you remember to start counting here (*point*). Put the green number, 3 (*point*), in your head. Now count up to the total, 5 (*point*). What color is the total in this problem?” (Student: Red.) “Yes, the total is written in red to tell us when to stop counting up. We put the green number in our head. Then, we count up with our fingers until we reach the red total, like this. When I count up, I use my fingers until I reach the total (*point*). I already have the green 3 in my head (*point to head*). I start counting up using the *next* number: 4 (*put up one finger*), 5 (*put up second finger*). How many fingers am I holding up?” (Student: 2.) “Yes, I have 2 fingers up. So,  $x$  equals 2. I know that 3 (*point*) + 2 (*write 2 above the x*) equals 5 (*point*). I know that  $x$  equals 2, and I write my answer like this:  $x = 2$ . You write the answer for me (*student writes*).” *If correct:* “That’s right! # + # equal #. So I write # in the blank. That tells me that the missing number,  $x$ , equals #.” *If incorrect:* “That’s not quite right. # (*point*) plus # equals #. So, the missing number is #. So I write # in the blank. That tells me that the missing number,  $x$ , equals #. Let’s try another one, but this time you’ll try to do the work” (*guide student using same procedures*).

Received December 27, 2007

Revision received April 21, 2008

Accepted April 21, 2008 ■

## Call for Nominations

The Publications and Communications (P&C) Board of the American Psychological Association has opened nominations for the editorships of **Developmental Psychology**, **Journal of Consulting and Clinical Psychology**, and **Psychological Review** for the years 2011–2016. Cynthia García Coll, PhD, Annette M. La Greca, PhD, and Keith Rayner, PhD, respectively, are the incumbent editors.

Candidates should be members of APA and should be available to start receiving manuscripts in early 2010 to prepare for issues published in 2011. Please note that the P&C Board encourages participation by members of underrepresented groups in the publication process and would particularly welcome such nominees. Self-nominations are also encouraged.

Search chairs have been appointed as follows:

- **Developmental Psychology**, Peter A. Ornstein, PhD, and Valerie Reyna, PhD
- **Journal of Consulting and Clinical Psychology**, Norman Abeles, PhD
- **Psychological Review**, David C. Funder, PhD, and Leah L. Light, PhD

Candidates should be nominated by accessing APA’s EditorQuest site on the Web. Using your Web browser, go to <http://editorquest.apa.org>. On the Home menu on the left, find “Guests.” Next, click on the link “Submit a Nomination,” enter your nominee’s information, and click “Submit.”

Prepared statements of one page or less in support of a nominee can also be submitted by e-mail to Emnet Tesfaye, P&C Board Search Liaison, at [etesfaye@apa.org](mailto:etesfaye@apa.org).

Deadline for accepting nominations is January 10, 2009, when reviews will begin.

# Catching Up or Falling Behind? Initial English Proficiency, Concentrated Poverty, and the Reading Growth of Language Minority Learners in the United States

Michael J. Kieffer  
Harvard University

This study contrasts growth trajectories in English reading for 2 groups of language minority (LM) learners—those who enter kindergarten with limited oral English proficiency and those who enter kindergarten proficient in oral English—with that of native English speakers. Fitting a multilevel model for change to longitudinal data on a nationally representative sample, this study examines students' growth trajectories from kindergarten through 5th grade. Three noteworthy findings emerge. First, LM learners entering kindergarten proficient in English have trajectories similar to those of native English speakers, but LM learners entering kindergarten with limited English have trajectories that diverge from those of native English speakers, yielding large differences in achievement by the 5th grade. Second, controlling for demographic risk factors, including socioeconomic status (SES), reduces the effect of initial English proficiency from large to moderate and yields differences that narrow over time. Finally, these differences depend on school poverty, with smaller differences evident among students in high-poverty schools than among students in low-poverty schools. Results emphasize the need for academic interventions for LM learners who enter school with limited English proficiency.

**Keywords:** reading development, second-language learners, achievement gaps, poverty, Early Childhood Longitudinal Study—Kindergarten Cohort

Cross-sectional studies have consistently found that students who come from homes in which a language other than English is spoken—referred to as language minority (LM) learners—have lower reading achievement in English than their native English-speaking peers (for a review, see August & Shanahan, 2006). However, few longitudinal studies have examined LM learners' growth trajectories in reading over time, particularly through the upper elementary school grades. Consequently, it is unclear to what extent these growth trajectories converge on, or diverge from, those of native English speakers.

The population of LM learners includes all students who have been exposed to a language other than English at home, whether they have limited English proficiency, are proficient in both languages, or are English dominant (August & Hakuta, 1997; August & Shanahan, 2006). At any given grade level, a subset of these learners will lack the English proficiency to gain full access to

mainstream instruction without support and thus be classified as English language learners or limited English proficient (LEP; August & Shanahan, 2006). Whereas LM learner status is conceptualized as a fixed characteristic that does not change over time, the English language learner and LEP labels are temporary classifications. Despite the recognition that these labels are temporary, much of the research that compares LM learners with native English speakers includes only those students classified as limited in proficiency at the time of the study. As a result, analyses of the achievement gaps between LM learners and native English speakers, such as those based on the National Assessment of Educational Progress results (e.g., National Center for Educational Statistics [NCES], 2007), present a biased picture of the achievement of the overall population of LM learners.

A preferable approach in developmental research has thus been to define the population by its time-invariant LM learner status (e.g. August & Hakuta, 1997; August & Shanahan, 2006). However, an overall comparison between the population of LM learners and the population of native English speakers may obscure qualitative differences within the heterogeneous group of LM learners. Of particular interest to developmental researchers is the extent to which differences among LM learners in initial English proficiency at the time of school entry predict differential growth in English reading. By sampling the entire population of LM learners and measuring their English language proficiency before formal schooling, a longitudinal analysis has the potential to examine the role of initial English language proficiency in later growth in English reading. Understanding this issue is fundamental to investigating how U.S. schools are meeting the instructional needs of this rapidly growing population.

---

The original collection of the ECLS-K data was funded by the National Center for Educational Statistics, U.S. Department of Education. This secondary analysis was funded in part by a Harvard University Presidential Fellowship and a Spencer Foundation Research Training Grant. I thank John B. Willett for his extensive methodological guidance and Nonie K. Lesaux and Catherine E. Snow for their valuable insights into the theoretical background and implications of this research. I also thank Kristen Bub, Mindy Munger, Young-Suk Kim, Gabrielle Rappolt-Schlichtmann, and Claire Vallotton for their helpful feedback on earlier versions of this article.

Correspondence concerning this article should be addressed to Michael J. Kieffer, Harvard Graduate School of Education, Harvard University, Larsen Hall 316, 14 Appian Way, Cambridge, MA 02128. E-mail: michael\_kieffer@gse.harvard.edu



## Examining English Reading Growth Among LM Learners

Although research in education and child development is often enhanced by the analysis of longitudinal data, analyses of growth over time are particularly essential to the study of LM learners. Cross-sectional comparisons of LM learners and native English speakers can suffer from cohort effects if the LM learners studied are only a subset of the population with an LEP label or if immigration patterns lead to differences between cohorts. Yet, the majority of studies that have modeled reading growth, as opposed to reading achievement, have been conducted with native English speakers (e.g., Compton, 2000; Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Foorman, Francis, Novy, & Liberman, 1991; O'Malley, Francis, Foorman, Fletcher, & Swank, 2002; Swanson & Jerman, 2007) or have used large data sets that included LM learners without examining LM learner status (e.g., Chatterji, 2006).

Among the few longitudinal studies that have been conducted with LM learners, the majority have only followed participants until second or third grade (e.g., Chiappe, Siegel, & Wade-Woolley, 2002; Gerber, Jimenez, Leafsteadt, Villaruz, Richards, & English, 2004; Manis, Lindsey, & Bailey, 2004; Stuart, 2004; Swanson, Saez, & Gerber, 2006; Verhoeven, 1990, 2000). Only two studies, to my knowledge, have followed participants beyond the primary grades (Lesaux, Rupp, & Siegel, 2007; Reese, Garnier, Gallimore, & Goldenberg, 2000). Although informative about the nature of reading development, these studies provide limited information about how growth trajectories in English reading might differ between native English speakers and LM learners in the United States. Lesaux et al. (2007) conducted their study in a Canadian city in which LM learners were evenly distributed across neighborhoods and socioeconomic levels, thus limiting their ability to generalize their findings to areas in the United States in which poverty and LM status are confounded. Reese et al. (2000) examined the English reading achievement of Spanish-speaking LM learners without the benefit of a native English-speaking comparison group.

### Theoretical Framework: Two Hypotheses for LM Learners' Growth in English Reading

The theoretical framework of this research acknowledges the developmental nature of reading and the possibility that the determinants of reading achievement change over time (RAND Reading Study Group, 2002). There is a general consensus that LM learners start school at elevated risk for reading difficulties, especially LM learners who enter kindergarten with limited English proficiency (August & Shanahan, 2006; Snow, Burns, & Griffin, 1998). However, there is much less understanding about how this risk affects reading growth over the course of the elementary school years. This study examines two hypotheses that describe the English reading development of LM learners relative to native English speakers: a *differential skills* hypothesis that posits that the growth trajectories in reading of the two groups will diverge over time and a *developmental lag* hypothesis that posits that the growth trajectories will converge over time, such that LM learners eventually catch up with native English speakers.

The differential skills hypothesis draws theoretical support from models of reading that posit that the relative contributions of component skills to reading achievement change over time, such as

Chall's (1983) stages of reading development and the RAND Reading Study Group's (2002) conceptualization of reading comprehension. Put simply, these models hypothesize that different skills come into play in the act of reading at the different grade levels. Whereas word-reading skills play the central role in reading achievement in kindergarten through third grade (e.g., Adams, 1990; Perfetti, 1988), oral language proficiency and higher order comprehension processes play greater roles in reading achievement in Grade 4 and beyond (e.g., Anderson & Freebody, 1983; RAND Reading Study Group, 2002; Vellutino, Scanlon, Small, & Tansman, 1991). Moreover, compared with early reading skills, proficient reading comprehension in later grades is more dependent on a wide array of experiences (RAND Reading Study Group, 2002) and on the knowledge accumulated from students' exposure to print (Cunningham & Stanovich, 1997).

The empirical research on the English reading development of LM learners generally supports this hypothesis. In a recent review, the National Literacy Panel on Language Minority Children and Youth found that LM learners can achieve levels equal to those of native English speakers at the word-reading tasks that predominate in the early grades, but they have low achievement on the reading comprehension tasks that become central in the upper elementary grades (Lesaux, 2006). In particular, studies using a longitudinal design have found word-reading development of LM learners on par with that of native English speakers (e.g., Chiappe et al., 2002; Lesaux et al., 2007; Lesaux & Siegel, 2003; Manis, Lindsey, & Bailey, 2004) or have found small differences in word reading between the two groups that disappear over time (Verhoeven, 1990, 2000). However, each of these studies has also found deficiencies in one or more aspects of LM learners' oral language skills in English—such as their vocabulary, syntactic awareness, and verbal working memory—that may predict future difficulties in English reading comprehension.

On this basis, it is expected that the growth trajectories in English reading of LM learners and native English speakers will diverge over the elementary school years, as reading becomes less about decoding words fluently and more about integrating sophisticated vocabulary knowledge, higher order thinking skills, and rich content knowledge to comprehend difficult texts. It is further expected that this divergence will affect LM learners with different initial levels of English oral proficiency differently. It is likely that students who enter kindergarten with limited oral English proficiency fall behind native English speakers at a faster rate than do LM learners who enter kindergarten more proficient in English. Indeed, one longitudinal study of a group of Spanish-speaking LM learners (Reese et al., 2000) detected a strong relationship between initial English proficiency at kindergarten entry and English reading achievement in seventh grade.

Alternately, the developmental lag hypothesis argues that given sufficient time and exposure to English, the reading growth trajectories of LM learners will converge on those of their native English-speaking peers. Currently, this hypothesis has limited theoretical or empirical support, although there is evidence that some LM learners can catch up with the academic achievement of their native English-speaking peers when given an extended window of opportunity, sometimes as long as 10 years (e.g., Collier, 1989; Hakuta, Butler, & Witt, 2000). Contrary to this, an assumption that LM learners will catch up with their classmates quickly is the basis for many current instructional programs that serve these learners. Transitional bilingual and English-as-a-second-language programs



typically provide services to students for 1 to 3 years, after which they are expected to be ready for mainstream classroom instruction in English (Zehler et al., 2002). Such programs presume that LM learners' English reading development converges with that of native English speakers once students have had the time to develop oral English proficiency.

Thus, testing whether the differential skills or developmental lag hypothesis best describes the growth trajectories in English reading of LM learners is a developmental question that informs how educational programs to serve these learners are conceptualized, designed, and implemented. This study evaluates these two hypotheses using a developmental framework to analyze longitudinal data on English reading achievement between kindergarten and fifth grade from a nationally representative sample.

### Poverty and LM Learner Status

Reading research has identified several factors—in addition to initial English proficiency—that are associated with greater risk of reading difficulties, including a low-income family background, being a student of color, and attending a school with a high concentration of poverty, a high concentration of students of color, or both (for a review, see Snow et al., 1998). Because LM learners in the United States are more likely to be poor, to be students of color, and to attend highly segregated schools (Capps et al., 2005; Cosentino de Cohen, Deterding, & Chu Clewell, 2005), it is important to disentangle the effects of these risk factors from the effects of LM learner status.

In particular, the negative effects that accrue from living in poverty and attending a high-poverty school likely contribute to the low achievement of LM learners. The confounding of SES and LM learner status in many analyses of reading achievement, such as those based on National Assessment of Educational Progress scores, makes it difficult to determine whether the apparent effects of LM learner status are instead a result of influences related to poverty. Indeed, there is some evidence that differences between LM learners and native English speakers are minimal when the two groups are matched on SES. For instance, in a study conducted in a Canadian school district in which the SES of LM learners was equal to that of native English speakers, Lesaux et al. (2007) found that LM learners demonstrated growth trajectories in English reading from kindergarten to fourth grade that were indistinguishable from those of native English speakers.<sup>1</sup>

Although socioeconomic differences may explain some of the achievement differences between LM learners and their native English-speaking peers in the United States, it is unlikely that LM learners' status has no effect independent of its relationship with poverty. As described above, the deficiencies in English oral language consistently found among many LM learners likely constrain their English reading development, and such deficiencies are likely to be less common or less pronounced among native English speakers experiencing similar levels of poverty but exposed to greater amounts of oral English at home. Moreover, the effects of poverty may be greater for LM learners than for their classmates if their limited English proficiency prevents them from accessing educational resources meant to ameliorate the effects of poverty, such as Title I supplementary educational services. A common argument among advocates for improving the education of LM learners is that these learners have less access to educational

resources than their native English-speaking peers, even within the same high-poverty schools (Gándara, Rumberger, Maxwell-Jolly, & Callahan, 2003).

### Research Questions

The present study is a secondary analysis of multiwave data on a nationally representative sample of U.S. elementary school children from the Early Childhood Longitudinal Study—Kindergarten Cohort (ECLS-K) dataset. The purpose of the study is to describe the differential English reading development trajectories of three subpopulations of students: (a) native English speakers, (b) LM learners who enter kindergarten with initially limited English proficiency (LM-iLEP),<sup>2</sup> and (c) LM learners who enter kindergarten with initially full English proficiency (LM-iFEP). The analyses describe these trajectories both as they exist on average in the national population and as they differ by children's demographic background and schooling context. Three sets of research questions guide this study:

1. Do growth trajectories in English reading of LM learners diverge from those of native English speakers over the elementary school years, thus supporting the differential skills hypothesis? Among LM learners, is this divergence greater for students who are LM-iLEP compared with those who are LM-iFEP?
2. Do growth trajectories in English reading of LM learners (LM-iLEP and LM-iFEP) converge with those of native English speakers from similar demographic backgrounds attending demographically similar schools over the elementary school years?
3. Does the effect of selected student and/or school demographic risk factors vary as a function of LM learner status, such that LM learners are more vulnerable to the effects of these risk factors?

### Method

#### *Dataset*

This study uses data from the ECLS-K (Tourangeau, Lê, & Nord, 2005), a study conducted by NCES that examines the complex roles of school, home, and individual characteristics on students' academic and behavioral development over the school years. In conducting the ECLS-K, NCES used a multistage probability sampling design to select a cohort of students that was

<sup>1</sup> However, it is also worth noting that the district in which this study was conducted had a strong early literacy program, including early intervention efforts for students at risk for reading difficulties; although not explicitly focused on LM learners, such instructional efforts may have contributed to the equal success of these learners.

<sup>2</sup> Educators and researchers (e.g., August & Hakuta, 1997) have expressed objections to the term *LEP students* as a negative label with connotations of a permanent deficit. I agree with this objection and use LM-iLEP in this article solely to describe an initial condition rather than to describe children and use the convention of people-first language by using "students who were LM-iLEP" rather than "iLEP students."



nationally representative of students entering kindergarten in the 1998–1999 school year; as such, it includes children from a wide variety of socioeconomic, ethnic, and linguistic backgrounds attending a wide variety of public, private, and parochial schools. As described in detail in Tourangeau et al. (2005), NCES took major steps to minimize attrition and preserve the representativeness of the sample over time. As the only nationally representative longitudinal study designed to include school-age LM students, this dataset is well suited to addressing the research questions.

The analytic sample includes 17,385 students from the ECLS-K study who had one or more reading score; this sample includes all but 180 of the students from the original sample, preserving its representativeness.<sup>3</sup> Approximately 13% of the sample reported speaking a non-English language at home; roughly half of these were students classified as LM-iLEP in the fall of kindergarten. Students were then assessed longitudinally a maximum of six times in the fall of kindergarten, spring of kindergarten, fall of first grade (for a random subsample of students), spring of first grade, spring of third grade, and spring of fifth grade. The analyses incorporated baseline demographic and language data obtained in the fall of kindergarten and multiple waves of English reading achievement data obtained from spring of kindergarten through the spring of fifth grade. ECLS-K methodological reports provide detailed information on the data collection procedures (Tourangeau et al., 2005).

## Measures

*English reading achievement.* The outcome of interest is children's overall English reading achievement. It was assessed with an individually administered, two-stage reading test assembled by a panel of content experts using items from published standardized tests (Pollack, Atkins-Burnett, Najarian, & Rock, 2005). Drawing on and extending the National Assessment of Educational Progress framework, the test assessed basic reading skills (familiarity with print, recognition of letters, recognition of phonemes, and decoding), vocabulary, and five types of reading comprehension skills (initial understanding, developing interpretation, personal reflection and response, and developing a critical stance). The analyses used the number-right true scores scaled using item response theory as the outcome, as these scores were constructed to be vertically equatable across occasions of measurement. As reported in Pollack et al. (2005), estimates of internal consistency reliability for these scores were high (.93 to .96, depending on wave). In field testing, the reading scores had a high correlation (.83) with scores from the construct validation instrument, the Woodcock-McGrew-Werder Mini-Battery of Achievement; this cross-test correlation was higher than the cross-content correlations with the other ECLS-K cognitive measures (.73 with math and .72 with science), providing evidence for the validity of inferences about students' reading achievement based on these scores. Two limitations of the ECLS-K measures are worth noting. First, the lack of a commonly used standardized instrument makes it difficult to make direct comparisons between these scores and scores reported in other studies. Second, the reading achievement of all LM learners was assessed in English only, preventing the examination of reading abilities in students' native languages.

*Time.* Time was specified as the child's age in months, measured at each wave and centered on each student's age at the

testing date in the spring of fifth grade (CAGE). Thus, the intercept in the hypothesized individual growth trajectory in reading represents the student's true "final" reading status in the spring of fifth grade. This final-status specification was selected on the basis of a substantive interest in whether LM learners ultimately catch up with native English speakers by the end of the elementary school years; with this specification, the main effects of the predictors can be interpreted easily as the ultimate differences in reading achievement associated with these variables. The downside of this specification is that the effects of predictors on final rate of change and acceleration are more difficult to interpret; readers are thus provided with fitted growth plots and corresponding effect sizes to ease interpretation of differences in the shape of growth trajectories. Because CAGE was centered on each student's age, a second time-invariant control predictor was also included in the statistical models to represent each student's deviation in months of age from the sample median final age as well as interactions between this predictor and the linear and quadratic terms for CAGE. Details on this and other variable specifications are available from me.

*English language proficiency group.* The predictor for Research Question 1 was the students' time-invariant initial classification into one of three English language proficiency groups, based on their LM learner status and oral English proficiency at school entry: (a) native English speakers, (b) LM learners who were LM-iLEP, and (c) LM learners who were LM-iFEP. First, children were classified as LM learners or native English speakers on the basis of parental report in the fall of kindergarten (for approximately 90% of cases), in the spring of kindergarten (for approximately 8% of cases), or in the spring of first grade (for fewer than 2% of cases). Second, LM learners were classified as either LM-iLEP or LM-iFEP on the basis of whether their scores on the English Pre-Language Assessment Scales (Pre-K LAS; Duncan & DeAvila, 1998) at school entry (i.e., in the fall of kindergarten for more than 90% of children) fell above or below a cut-score established by the authors of the Pre-K LAS.<sup>4</sup> Note that this construct is a time-invariant classification, referring to students' LM learner status (i.e., LM or native English) and initial English language proficiency and should not be confused with students' concurrent English language proficiency, which would be expected to change over time. This categorical predictor consisted of two dummy variables representing LM-iLEP status and LM-iFEP status, with native English speakers specified as the

<sup>3</sup> These students were excluded because they did not have reading scores at any time point. Some were students with severe disabilities who were thereby excluded from all the cognitive measures, whereas others were LM learners who were excluded from testing in early rounds because of limited English proficiency and were then lost to follow-up in the third- and fifth-grade rounds (during which all students were tested regardless of English proficiency). Although it is therefore likely that these 180 students differ systematically from the original sample of 17,565, their small number indicates only a very small threat to the validity of generalizations to the population of U.S. school children who entered kindergarten in 1998.

<sup>4</sup> It is worth noting that the Spanish oral language proficiency of Spanish-speaking LM learners was also assessed at school entry. However, these scores are not included in the present analyses for two reasons. First, comparable scores were not available for students speaking other native languages. Second, the role of native language proficiency in second-language reading achievement is beyond the scope of the present investigation.

reference category (i.e., coded with a 0 for both LM-iLEP status and LM-iFEP status). Thus, the intercept can be interpreted as the predicted value for native English speakers.

*School-level and student-level demographic variables.* The student-level demographic variables included were native language background (Spanish or another language), ethnicity, and SES (represented by an ECLS-created composite of parents' education, occupation, and household income). School-level demographic variables included concentration of poverty (measured as the percentage of students receiving free or reduced-price lunch), concentration of students of color, and concentration of students with a school-based designation of LEP, each based on administrator reports. All student-level and school-level control predictors were declared time invariant; their values were fixed to the average value of the child's available values across the five occasions of data collection, with the exceptions of ethnicity and native language.

Preliminary analyses indicated that, as hypothesized, LM learner status was largely confounded with demographic risk factors. Table 1 displays the sample means of the demographic variables included, by English language proficiency group, along with the results of *t* tests of population equivalences in mean among the groups. The column on the far right indicates that when compared with native English speakers, LM learners were more likely to be Latino or Asian/Pacific Islander and tended to have lower SES (by approximately 0.60 standard deviation). On average, LM learners also attended schools with much higher concentrations of poverty and higher concentrations of both students of color and students designated as LEP than did native English speakers. It is worth noting within the population of LM learners the demographic differences between those who entered kindergarten proficient in English and those who entered with limited proficiency in English. The third column indicates that compared

with students in the LM-iLEP group, students in the LM-iFEP group were more likely to be Asian or Pacific Islander and were less likely to be Latino and tended to have higher SES (by more than .50 standard deviation). On average, students who were LM-iFEP attended schools lower in poverty, lower in concentrations of students of color, and lower in concentrations of students designated as LEP than did students who were LM-iLEP.

*Late qualifying.* There was variation in the age at which LM learners gained enough English proficiency to be administered the reading assessment. Students who did not pass the Pre-K LAS at each wave in kindergarten and first grade were excluded from taking the reading test, whereas those LM learners who passed the Pre-K LAS took the reading test in that round and in all subsequent rounds. In third and fifth grade, all students were assessed on the reading test regardless of English language proficiency. Thus, a simple estimate of the effect of LM-iLEP status may be positively biased by the systematic exclusion of the group of students who were the slowest to acquire English (and therefore also likely to be low in reading). Table 2 displays the frequencies and sample proportions of LM learners who were LM-iLEP and excluded from testing and the proportion who qualified at each wave of data collection. The exclusion of large proportions of students who were LM-iLEP in the early waves suggests that the potential bias may be substantial. To assess the effect of the exclusion of students who were LM-iLEP from reading testing in the early waves because of limited English proficiency, time-invariant dummy control predictors were included in some of the statistical models to identify the periods in which students were late-qualified for the reading test after being previously excluded (iLEP-QUALTIME2, iLEP-QUALTIME3, and iLEP-QUALTIME4).

Table 1

*Sample Means and Differences Between the Means of Selected Child- and School-Level Demographic Characteristics for Language Minority (LM) Learners Who Entered Kindergarten With Initially Full English Proficiency in English (LM-iFEP), LM Learners Who Entered Kindergarten With Initially Limited English Proficiency (LM-iLEP), and Native English Speakers*

Variable	LM-iFEP ( <i>n</i> = 746)	LM-iLEP ( <i>n</i> = 1,134)	Difference between LM-iFEP & LM-iLEP <sup>a</sup>	Native English ( <i>n</i> = 15,362)	Difference between native English & all LM <sup>a</sup>
<b>Child variables</b>					
Ethnicity					
White	0.133	0.020	0.113*	0.573	0.508***
Latino	0.474	0.727	-0.253***	0.113	-0.527***
African American	0.041	0.002	0.039***	0.159	0.152***
Asian/Pacific Islander	0.339	0.245	0.094***	0.046	-0.254***
Native American	0.000	0.002	-0.002	0.020	0.020***
Multiracial or not specified	0.006	0.001	0.005*	0.029	0.026***
Child socioeconomic status	-0.220	-0.766	0.545***	0.040	0.579***
Spanish speaking	0.454	0.724	-0.270***	0.020	-0.636***
<b>School variables</b>					
Concentration of poverty	55.921	70.350	-14.429***	44.287	-20.307***
Concentration of students of color	60.107	75.776	-15.670***	34.262	-36.261***
Concentration of students who are LEP	19.636	36.072	-16.436***	5.319	-24.772***

<sup>a</sup> Positive differences favor students who were LM-iFEP in the middle column and native English speakers in the right column. Approximate *p* values accompanying mean differences are derived from *t* tests of differences between pairs of identified groups. LM-iLEP = language minority learners who entered kindergarten with initially limited English proficiency; LM-iFEP = language minority learners who entered kindergarten with initially full English proficiency.

\*\*\* *p* < .001.



Table 2  
Sample Frequencies and Proportions of Language Minority (LM) Learners Who Both Entered Kindergarten With Initially Limited English Proficiency (iLEP) and Qualified or Were Systematically Excluded From Reading Assessment in Each Wave of Data Collection

Wave	Cumulative frequency of students who were LM-iLEP & qualified	Proportion of students who were LM-iLEP and late qualified	Proportion of students who were LM-iLEP who were excluded
Time 0 (Fall, Kindergarten)	0	—	1.00
Time 1 (Spring, Kindergarten)	577	—	0.60
Time 2 (Fall, 1st grade)	646	0.05	0.55 <sup>a</sup>
Time 3 (Spring, 1st grade)	1,087	0.33	0.25
Time 4 (Spring, 3rd grade)	1,408	0.25	0.00
Time 5 (Spring, 5th grade)	1,408	—	0.00

<sup>a</sup> This is an extrapolation based on the proportion of the students who were not qualified to take the test in this round had they participated in it, not the actual proportion excluded in this subsample. The frequency of students who qualified does not match the frequency of reading scores for students who were LM-iLEP because many students theoretically qualified for the reading test but did not participate in later waves of the study for other reasons.

Data Analysis

To address the research questions, a taxonomy of multilevel models for change was fitted in the person-period dataset that contained the longitudinal data on all sampled children, using STATA XT MIXED with maximum likelihood estimation. The multilevel model for change provides a powerful tool for addressing questions concerning systematic interindividual differences in change over time in longitudinal data (Singer & Willett, 2003). Preliminary inspection of empirical growth plots of each child's reading achievement as a function of age suggested that a quadratic growth specification was the most appropriate for representing the individual growth trajectories in reading. Thus, the multilevel model for change hypothesized to address the first research question, expressed in composite form, was

$$READING_{ij} = \left( \begin{array}{l} \gamma_{00} + \gamma_{10}CAGE_{ij} \\ + \gamma_{20}CAGE_{ij}^2 + \gamma_{01}CFINALAGE_i \\ + \gamma_{11}CFINALAGE_i \times CAGE_{ij} \\ + \gamma_{21}CFINALAGE_i \times CAGE_{ij}^2 \\ + \gamma_{02}ILEP_i + \gamma_{12}ILEP \times CAGE_{ij} \\ + \gamma_{22}ILEP \times CAGE_{ij}^2 \\ + \gamma_{03}IFEP_i + \gamma_{13}IFEP \times CAGE_{ij} \\ + \gamma_{23}IFEP \times CAGE_{ij}^2 \end{array} \right) + \zeta_{0i} + \zeta_{1i} \times CAGE + \varepsilon_{ij}$$

where  $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$  and  $\begin{bmatrix} \zeta_{0i} \\ \zeta_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix}\right)$

where parameter  $\gamma_{00}$  represented the population average true final (spring of 5th grade) status, parameter  $\gamma_{10}$  represented the population average true final instantaneous slope, and parameter  $\gamma_{20}$  represented the population average true acceleration. Parameters  $\gamma_{01}$ ,  $\gamma_{11}$ , and  $\gamma_{21}$  represented the effects of variation in students' age at the final date of testing on the final status, final instantaneous slope, and acceleration. Parameters  $\gamma_{02}$  and  $\gamma_{03}$  represented the effects on final achievement of LM-iLEP and LM-iFEP status, respectively. Parameters  $\gamma_{12}$  and  $\gamma_{13}$  represented the effects on final rate of growth of LM-iLEP and LM-iFEP status, respectively. Parameters  $\gamma_{22}$  and  $\gamma_{23}$  represented the effects on acceleration of LM-iLEP and LM-iFEP status, respectively. The random effect  $\varepsilon_{ij}$

was a Level 1 residual for child  $i$  at time  $j$  and was assumed to be drawn from a normal distribution with mean of 0 and unknown variance  $\sigma_\varepsilon^2$ . Random effects  $\zeta_{0i}$  and  $\zeta_{1i}$  represented Level 2 residuals for intercept and slope, respectively. They were hypothesized to be drawn from a multivariate normal distribution with a mean vector of zero, unknown variances  $\sigma_0^2$  and  $\sigma_1^2$ , and unknown covariance  $\sigma_{01}$ . Attempts to include a random effect for the acceleration term created convergence problems, leading to the decision to fix this effect across individuals.

The hypothesized multilevel models to address the second and third research questions were similar to this model, with the addition of terms representing the effects of the student-level and school-level demographic variables described above on final status, final rate of growth, and acceleration, as well as the interactions between these variables and LM-iLEP and LM-iFEP status. Following the recommendation of Singer and Willett (2003), the likelihood ratio test was the primary index used to compare the goodness of fit of successive models. The Akaike information criterion and Bayesian information criterion were also estimated and compared; changes in both of these indices corroborated the conclusion of the likelihood ratio test in all but one case, in which the conclusion indicated by the likelihood ratio test and Akaike information criterion was selected over that indicated by the Bayesian information criterion.

Missing Data

Some data were missing in the ECLS dataset both at Level 1 (within child) and at Level 2 (between child), with the most substantial data missing in the school-level demographic variables at each wave and in the outcome at the fifth-grade wave. Several strategies were used to make use of all legitimate, available information without imputing values or excluding cases (either of which would have imposed more stringent assumptions on the model). First, as recommended by Singer and Willett (2003), any child contributing one or more waves of reading score to the person-period dataset was included in the analyses and contributed the available multiwave information they provided to the Level 1 estimation, whereas missing waves of reading scores were simply not included in the person-period dataset or analyses. Second, the

systematic missingness in Level 1 data due to the exclusion of students with limited English proficiency from the reading test in early waves was taken into account explicitly through the inclusion of the late-qualifying variables described above. Third, by specifying the control variables as time invariant and fixing these to the average of the available values across the five waves of data collection, any student who had survey data from one or more waves was able to have an empirically based value for these variables. Following this decision, only 1% of students lacked data on a student-level control variable, and 18% of students lacked data on a school-level control variable. Fourth, Cohen's (2003) model-based approach was used to include these remaining children; following this approach, arbitrary values ( $-9$ ) were assigned to any missing value for any control predictor, and additional dummy predictors were created and included to accompany these predictors to indicate which values were missing and replaced by placeholders. Resulting parameter estimates were thus the product of all available nonmissing information, and values did not need to be imputed on the basis of some specified set of predictors.

These four strategies allowed for the inclusion of 17,385 students, constituting more than 99% of the original ECLS-K participants, and 62,992 observations, constituting 85% of the observations (i.e., reading scores) that would be possible if there was no attrition or missing waves of outcome data. Supplemental analyses indicated that the absence of observations was in large because of the late qualifying of LM learners described above and to students moving between elementary schools (details are available from me). Regarding the latter concern, NCES followed only a subsample of 50% of students who moved between schools, selected at random.<sup>5</sup> This random selection ensured that there were no systematic differences between the movers with data and those who were missing data.

### *Accounting for the Complex Sampling Design of the ECLS-K*

In generalizing to the population of U.S. students who entered kindergarten in 1998, two aspects of the ECLS-K complex sample survey design warranted attention. First, to ensure adequate sample size in making comparisons, ECLS-K oversampled selected groups of children. In the base year, this included children who were Asian or Pacific Islander and children who attended private schools. In subsequent years, LM learners who changed schools were also oversampled, as compared with native English speakers who changed schools. Given the difficulties of incorporating details of the complex survey sample design into multilevel analyses with current statistical software, oversampling was accounted for by including predictors in the models that represented the dimensions on which the oversampling occurred (i.e., ethnicity, school type, and language minority learner status).

Second, ECLS-K used a multistage probability sampling design in which the primary sampling units (PSUs) were geographic areas consisting of counties or groups of counties, the second-stage units were schools, and the third-stage units were students within schools (Tourangeau et al., 2005). Accounting for this multistage sampling design while simultaneously fitting an appropriate multilevel model for change provided technical challenges that are not yet fully met by the statistical software currently available. The

STATA software package currently allows the fitting of multilevel models with relatively simple error structure while stipulating the PSUs (using the XTREG command and generalized least squares estimation) or the fitting of multilevel models with more complicated error structures without stipulating the PSUs (using the XTMIXED command and maximum likelihood estimation), but not both at the same time. Therefore, two parallel sets of analyses were conducted, in which first a taxonomy of multilevel models for change was fitted with the error structure described above but without accounting for the PSUs, using XTMIXED, and then selected models from this taxonomy were refitted using a simplified error structure (i.e., a multilevel model with random intercepts only), but accounting for the PSUs, using XTREG.

## Results

The findings for each of the three research questions are presented below. The first section presents results for models that compared the growth trajectories in English reading of LM learners and native English speakers without taking into account SES, race, or school demographics. This analysis provided the description of average achievement differences as they actually existed in U.S. schools. The second section presents results for models that took into account the demographic predictors of child ethnicity, child SES, school concentration of poverty, and school concentration of students of color. The magnitudes of these predictors were evaluated, along with the impact of including these predictors on the magnitudes of the effects of LM learner status. The third section presents results for a final model that included these effects and an interaction between LM learner status and school-level poverty.

For ease of interpretation, fitted growth trajectories were estimated for prototypical cases and displayed in graphical form, and effect sizes were estimated for differences in the elevation of these trajectories between LM learners and native English speakers at kindergarten, third grade, and fifth grade. Using effect sizes provides a standardized metric for assessing the magnitude of these differences and allows one to compare them against Cohen's (1992) conventions in which an effect size of .2 indicates a small effect, .5 indicates a medium effect, and .7 indicates a large effect. Comparing effect sizes is also preferable to evaluating the statistical significance of the parameter estimates alone; given the large sample size and its high statistical power, even very small effects with little practical importance can be statistically significant. Additionally, it is worth noting that all models were fitted using the complete dataset. Thus, the population described included students from diverse ethnic, linguistic, and socioeconomic backgrounds who attended U.S. schools. For the purposes of displaying fitted growth trajectories, all control variables had to be temporarily set to some prototypical values, thus the trajectories displayed were for Latino students with average SES who attended high- and low-poverty public schools with an average concentration of students of color. However, readers should bear in mind that analyses included data from all students and thus findings are not limited to this subpopulation.

<sup>5</sup> Language minority learners who moved schools were followed at a 100% rate.



Research Question 1: Many LM Learners Are Falling Far Behind the National Average

The first research question investigated the differential growth trajectories in English reading of LM learners and native English speakers. The findings indicated that LM learners who entered school with limited English proficiency had growth trajectories in reading that unfolded at substantially lower elevations than those of native English speakers throughout the elementary school years, but that LM learners who entered school proficient in English had growth trajectories in reading that were largely indistinguishable from those of native English speakers.

Table 3 displays a taxonomy of fitted multilevel models for change obtained in these analyses. The unconditional growth model, labeled *Model 1* in the table, indicated that as expected, a quadratic specification of reading achievement as a function of age described the curvilinear shape of reading development over time. On average, children’s reading achievement tended to increase with age ( $\hat{\gamma}_{10} = 0.477, p < .001$ ), but at a decelerating rate ( $\hat{\gamma}_{20} = -0.020, p < .001$ ).

Model 2 assessed the impact of LM-iLEP and LM-iFEP status on students’ fifth-grade reading status, final rate of growth, and acceleration in reading achievement, thus provid-

Table 3  
Taxonomy of Fitted Multilevel Models for Change in Which Reading Achievement Is a Quadratic Function of the Child’s Age, English Language Proficiency Group (LM-iLEP and LM-iFEP Compared With the Reference Category, Native English) and Late Qualification for the Reading Assessment Because of Limited English Proficiency (N = 17,385)

Variable	Symbol	Model 1: Unconditional growth model	Model 2: Incl. language group	Model 3: Incl. language + late qualification
Fixed effects				
Final status, $\pi_{0i}$				
Intercept	$\gamma_{00}$	138.863***	140.234***	140.320***
CFINALAGE	$\gamma_{01}$	0.125**	0.052	0.049
LM-iLEP	$\gamma_{02}$		-19.571***	-10.059***
LM-iFEP	$\gamma_{03}$		-0.637	-0.709
iLEP—QUALTIME2	$\gamma_{04}$			-5.056
iLEP—QUALTIME3	$\gamma_{05}$			-11.290***
iLEP—QUALTIME4	$\gamma_{06}$			-20.780***
Final rate of change, $\pi_{1i}$				
Intercept	$\gamma_{10}$	0.477***	0.467***	0.468***
CFINALAGE	$\gamma_{11}$	-0.032***	-0.031***	-0.031***
LM-iLEP	$\gamma_{12}$		0.194***	0.194***
LM-iFEP	$\gamma_{13}$		0.012	0.012
iLEP—QUALTIME2	$\gamma_{14}$			-0.205
iLEP—QUALTIME3	$\gamma_{15}$			-0.001
iLEP—QUALTIME4	$\gamma_{16}$			0.146
Acceleration, $\pi_{2i}$				
Intercept	$\gamma_{20}$	-0.020***	-0.020***	-0.020***
CFINALAGE	$\gamma_{21}$	-0.0002***	-0.0002***	-0.0002***
LM-iLEP	$\gamma_{22}$		0.006***	0.005***
LM-iFEP	$\gamma_{23}$		-0.0004	0.0004
iLEP—QUALTIME2	$\gamma_{24}$			-0.005*
iLEP—QUALTIME3	$\gamma_{25}$			0.002
iLEP—QUALTIME4	$\gamma_{26}$			0.006**
Variance components				
Level 1				
Within-child	$\sigma_{\epsilon}^2$	101.059***	100.721***	100.642***
Level 2				
In intercept	$\sigma_0^2$	677.078***	643.948***	636.446***
In rate of change	$\sigma_1^2$	0.084***	0.082***	0.082***
Covariance	$\sigma_{12}$	5.939***	5.673***	5.607***
Goodness of fit				
Deviance		524,559.20	523,699.64	523,501.74
Akaike information criterion		524,579.20	523,737.60	523,557.70
Bayesian information criterion		524,669.70	523,909.60	523,811.20
Likelihood ratio tests				
$H_0$				
$\gamma_{02} = \gamma_{03} = \gamma_{12} = \gamma_{13} = \gamma_{22} = \gamma_{23} = 0$			859.56***	
$\gamma_{04} = \gamma_{05} = \gamma_{06} = \gamma_{14} = \gamma_{15} = \gamma_{16} = \gamma_{24} = \gamma_{25} = \gamma_{26} = 0$				197.90***

Note. LM = language minority; LM-iLEP = LM learners entered kindergarten with initially limited English proficiency; LM-iFEP = LM learners who entered kindergarten with initially full English proficiency; CFINALAGE = ; QUALTIME2 = qualified at Time 2 to take the reading test; QUALTIME3 = qualified at Time 3 to take the reading test; QUALTIME4 = qualified at Time 4 to take the reading test.  
\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

ing a preliminary answer to the first research question. In this model, and in later models, the reference category was native English speakers; thus, the coefficients for the effects of LM-iLEP and LM-iFEP can be interpreted as the difference between these groups and native English speakers. Compared with native speakers, students who were LM-iLEP had a substantially lower fifth-grade status ( $\hat{\gamma}_{02} = -19.571, p < .0001$ ), a higher final rate of growth ( $\hat{\gamma}_{12} = 0.194, p < .0001$ ), and a rate of acceleration that was slightly but significantly higher ( $\hat{\gamma}_{22} = 0.006, p < .0001$ ). Students who were LM-iFEP did not differ from native English speakers in their fifth-grade status ( $p = .562$ ), their final rate of growth ( $p = .749$ ), or their rate of acceleration ( $p = .454$ ). When the complex survey design was taken into account by refitting the multilevel models with a simplified error structure and accounting for the PSUs using GLS estimation, each of these findings was found to be robust—the magnitudes of the effects were essentially the same, and there were no differences in the statistical inference.

As noted earlier, these estimates of the effect of LM-iLEP may be positively biased by the systematic exclusion of many students with low English proficiency from taking the reading test in Waves 1 through 3. As anticipated, accounting for the exclusion of these

students' scores in the early waves did in fact have an impact on the magnitudes of these estimates of the differences in trajectories. Model 3 in Table 3 displays a fitted multilevel model for change that included the set of dichotomous control predictors that distinguished the late qualifying of students who were LM-iLEP in each round (predictors iLEP-QUALTIME2, iLEP-QUALTIME3, and iLEP-QUALTIME4). In this model, likelihood ratio tests confirmed that the set of predictors describing late qualifying simultaneously predicted final status ( $p < .0001$ ), final rate of change ( $p < .0001$ ), and acceleration ( $p = .0082$ ). Given the inclusion of these variables, readers are cautioned in interpreting the parameter estimate for LM-iLEP in Models 3 to 6; in each of these models, this estimate refers to the effect of LM-iLEP status for only those students who were LM-iLEP and qualified for reading testing by the spring of kindergarten. A better interpretation of the effect of LM-iLEP status can be gained by looking across the parameter estimates for LM-iLEP, iLEP-QUALTIME2, iLEP-QUALTIME3, and iLEP-QUALTIME4 or by interpreting the fitted growth trajectories described below.

To provide insight into these findings, Figure 1 displays fitted growth trajectories in reading for native English speakers, for students who were LM-iFEP, and for students who were LM-iLEP

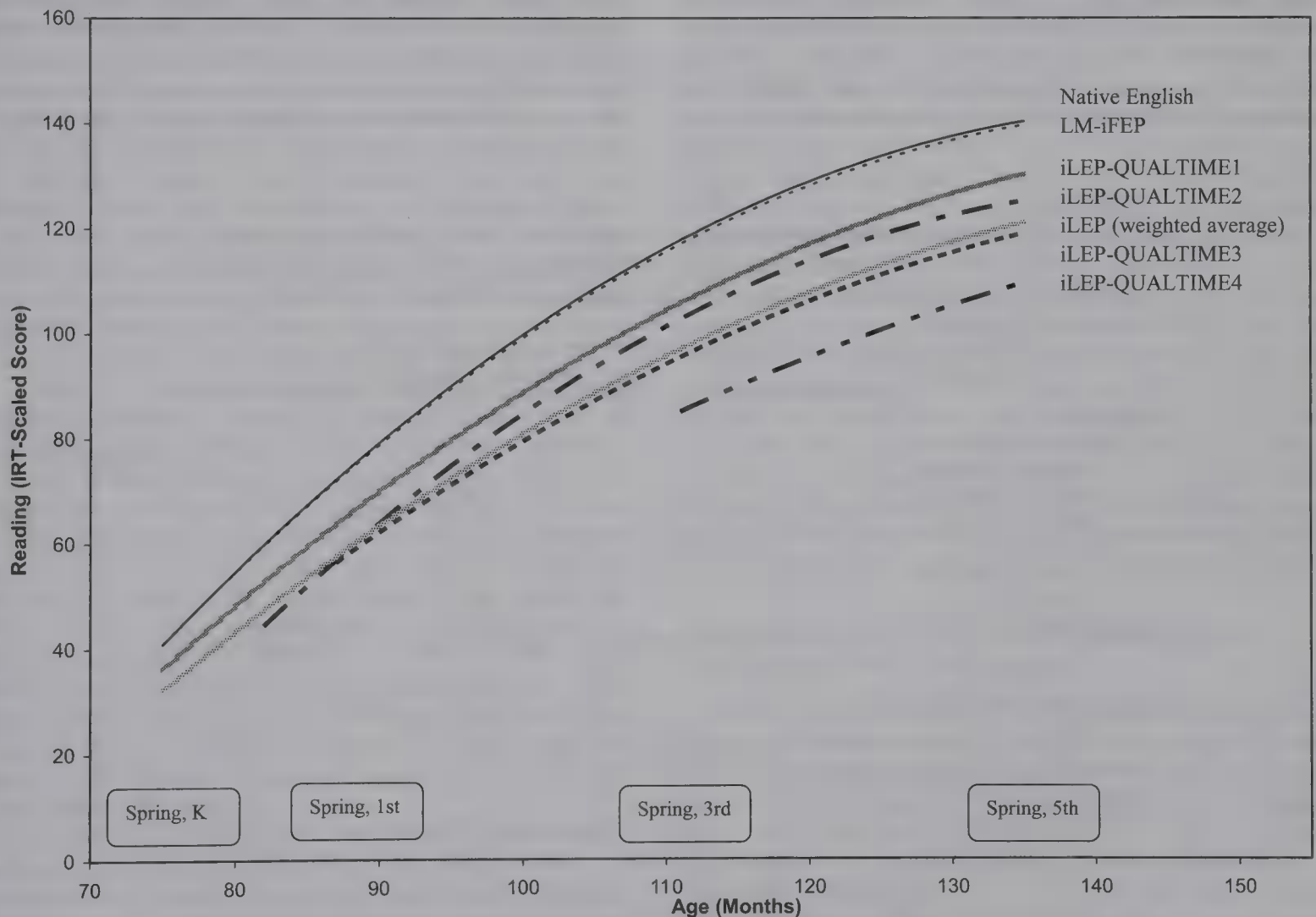


Figure 1. Fitted growth trajectories in reading for native English speakers, language minority learners who entered kindergarten with initially full English proficiency (LM-iFEP), and language minority learners who entered kindergarten with initially limited English proficiency (LM-iLEP) and late qualified for the reading test in each round ( $N = 17,385$ ). IRT = item response theory.



and qualified to take the test at each round. First, comparing the solid and dotted fitted trajectories at the top of the figure, notice that the predicted reading growths of native English speakers and of LM learners who entered kindergarten proficient in English are largely indistinguishable across the elementary school years.

Second, comparing the solid fitted trajectory with the thickly dashed fitted trajectories below reveals that LM learners who entered kindergarten with limited English proficiency had large and persistent gaps in English reading achievement compared with native English speakers throughout the elementary school years. The gray dotted line represents a weighted average of the growth trajectories in reading of the LM-iLEP groups that qualified for the reading test at different times; as such, it provides a single summary of what the average reading trajectory for students who were LM-iLEP could have looked like if all students had been included in the reading assessment from the spring of kindergarten. Comparing this weighted average fitted trajectory with the solid line that represents the average fitted reading trajectory for native English speakers reveals that the predicted gap in achievement between these groups increased between kindergarten and third grade, but then remained relatively stable between third and fifth grade.

Third, a comparison among the four different dashed fitted growth trajectories in reading shows that the excluded students were systematically lower in English reading achievement than other students who were LM-iLEP but assessed at every wave. The elevations of the fitted trajectories for each of the late-qualifying groups fell beneath those of the students who were LM-iLEP but qualified at Time 1 to take the reading test, substantially so in the

case of students who qualified at Time 3 or Time 4. The lowest achieving group consisted of those LM learners who did not acquire enough English by the end of first grade to take the reading test, as represented by the lowest dashed line on Figure 1; their predicted achievement in the spring of fifth grade was more than 29 item response theory scale score points (or approximately 1.25 standard deviation) below that of native English speakers. These students, who made up roughly 25% of the students who were LM-iLEP, may have been qualitatively different from the other two groups of late qualifiers. Technically, they never met the standard for English proficiency to take the reading test in first grade, but rather “qualified” for the reading test by virtue of an administrative decision to test all students in reading in third grade without first screening their English proficiency; although it is likely that many of them would have met the English proficiency standard in third grade, these data do not allow us to know what proportion of them would not have or to model how this might have predicted their growth.

Overall, when compared with native English speakers, students who were LM-iLEP, whether included or not in testing in early rounds, demonstrated fitted growth trajectories in English reading that were substantially lower in elevation and that diverged substantially between kindergarten and third grade. These differences persisted into fifth grade, but did not increase. To interpret the magnitude of these effects, the top half of Table 4 provides effect sizes for the fitted effects of LM-iLEP and LM-iFEP status (i.e., the vertical separation between the trajectories of these learners and those of native English speakers) for the spring of kindergar-

Table 4  
*Predicted Effect Sizes in Reading Achievement for LM Learners Who Entered Kindergarten With Initially Full English Proficiency (iFEP) and LM Learners Who Entered Kindergarten With Initially Limited English Proficiency (iLEP) as Compared With Native English Speakers (N = 17,385)*

Model		Effect size (Cohen's <i>d</i> )		
Model 2				
Effect of LM-iFEP				
Spring, Kindergarten			0.01	
Spring, 3rd grade			-0.03	
Spring, 5th grade			-0.03	
Effect of LM-iLEP				
Spring, Kindergarten			-0.71	
Spring, 3rd grade			-0.83	
Spring, 5th grade			-0.84	
Model 3	Highest estimate	Lowest estimate	Weighted average	
Effect of LM-iLEP				
Spring, Kindergarten		—	-0.34	-0.63
Spring, 3rd grade		-1.30	-0.48	-0.82
Spring, 5th grade		-1.33	-0.43	-0.83
Model 6 (in medium-poverty schools)				
Effect of LM-iLEP				
Spring, Kindergarten		—	-0.32	-0.57
Spring, 3rd grade		-0.77	-0.21	-0.44
Spring, 5th grade		-0.74	-0.11	-0.39

*Note.* In each model, the highest estimate of the gap between native speakers and the LM-iLEP group comes from those students who failed to pass the English proficiency test in the spring of first grade and therefore did not take the reading test until third grade, whereas the lowest estimate of the gap comes from those who passed the English proficiency test by the spring of kindergarten. The estimates from Model 6 were fitted for Latino children of average socioeconomic status who attended public schools of average concentration of poverty and average concentration of students of color.

ten, the spring of third grade, and the spring of fifth grade, based on Model 2. Because the difference in elevation between the trajectories of native English speakers and students who were LM-iLEP varied by when they qualified for the reading test, Table 4 also provides a range of estimated effect sizes for LM-iLEP. These include the lowest estimates (for students who were LM-iLEP and qualified at Time 1), the highest estimates (for students who were LM-iLEP and qualified at Time 4), and an average of estimated effect sizes, weighted by the proportions of students who were LM-iLEP and qualified in each round.

Consistent with the previous interpretation of the fitted growth trajectories, comparison of these effect sizes indicated on one hand that the English reading achievement of LM learners who entered kindergarten proficient in English differed slightly, if at all, from that of native English speakers in the elementary school grades. On the other hand, LM learners who entered kindergarten with limited English proficiency had medium-sized deficiencies in English reading achievement at the end of kindergarten that became large in the early elementary years and maintained thereafter.

### *Research Question 2: LM Learners Are Catching Up to Their Classmates*

My second research question asked whether the detected pattern of widening gaps in English reading achievement holds when LM learners are compared with native English speakers from similar demographic backgrounds who attended demographically similar schools. The findings indicate that controlling for these demographic characteristics led to fitted growth trajectories in reading for LM learners and native English speakers that were much more similar than those fitted without taking the demographic characteristics into account. Controlling for demographic factors also yielded fitted growth trajectories for the three groups that converged over the elementary school years. Table 5 lists three fitted multilevel models for change that illustrate the key findings concerning the effects of these demographic characteristics.

In Model 4, child-level demographic controls were statistically significant predictors of reading achievement. As one would expect, both children's ethnicity and SES affected the final status, final rate of change, and rate of acceleration of their fitted growth trajectories in reading (all  $p$ s < .05), with students of color and students from lower socioeconomic backgrounds having predicted growth trajectories unfolding at a substantially lower elevation. However, no feature of the growth trajectory—final status, final rate of change, and acceleration—was affected by whether the child spoke Spanish at home (all  $p$ s > .1540), once the child's ethnicity and SES had been taken into account. Although we might have expected that the effect of English language proficiency group would have differed for students of different ethnicities, level of SES, or native language, no significant interactions were found among these predictors.

In Model 5, the school-level demographic characteristics—concentration of poverty, concentration of students of color, and concentration of students designated as LEP—were investigated. Among these school-level predictors, concentration of poverty had the largest effects on students' growth trajectories in reading, specifically a substantial negative impact on predicted fifth-grade status (−.09 standard deviation for every additional 25 percentage points of students receiving free or reduced-price lunch;  $p$  < .001)

and statistically significant, but small, positive effects on final rate of growth ( $p$  = .005) and rate of acceleration ( $p$  < .001). Once the school's concentration of poverty was taken into account, the concentration of students of color did not have a significant effect on final status ( $p$  = .085), but did have significant, if small, positive impacts on the final rate of change ( $p$  = .012) and acceleration ( $p$  < .001). The school's concentration of students designated as LEP did not have a significant impact on any aspect of students' growth trajectories once other school demographics were taken into account.

As shown in Table 5, the estimated effects of LM-iLEP and LM-iFEP status on the fitted growth trajectories in reading were changed by the inclusion of the student-level and school-level demographics. The parameter estimate for the effect of LM-iFEP on final status became positive and statistically significant; LM learners who entered school proficient in English were predicted to be .13 standard deviation higher in reading by fifth grade compared to native English speakers of the same ethnicity and SES who attended schools with the same concentration of poverty and students of color ( $p$  = .004). The coefficient for the effect of LM-iLEP on final status was no longer significant, indicating that LM learners who entered school with limited English and qualified for the reading test by the spring of kindergarten had fifth-grade reading achievement that was no different from that of native English speakers with the same demographic risk factors. However, the coefficients for the effect of iLEP-QUALTIME3 and iLEP-QUALTIME4 on final status continued to be negative, large, and significant after accounting for school demographics, indicating that students who were LM-iLEP and excluded from early reading testing (making up 55% of all students who were LM-iLEP) had fifth-grade reading achievement that was substantially lower than that of native English speakers with similar demographic risk factors.

In sum, the findings for Research Question 2 indicated that controlling for the confounding influence of demographic risk factors led to a substantial change in the observed achievement differences between LM learners and native English speakers. LM learners who entered school proficient in English had growth trajectories that were higher in elevation than those of their demographically similar peers who attended similar schools, whereas LM learners who entered school with limited English had trajectories that were only moderately lower in elevation than those of their demographically similar peers who attended similar schools.

### *Research Question 3: The Interaction Between LM Learner Status and School Poverty*

The third research question focused on the potential moderating impact of LM learner status on the effects of demographic risk factors on English reading achievement. To address this question, two-way statistical interactions between variables for LM learner status and the demographic predictors were included in the multilevel models for change. Among the several two-way interactions tested, only the interaction between school poverty and LM learner status had a statistically significant effect on reading achievement. In the final model, Model 6, the interaction between LM learner status and school poverty had a significant effect on final status ( $p$  = .0331), final rate of change ( $p$  = .0319), and acceleration ( $p$  = .0252). This finding indicated that among students of



Table 5  
*Taxonomy of Fitted Multilevel Models for Change in Which Reading Achievement Is a Function of the Child's Age, English Language Proficiency Group (Native English, LM-iLEP, or LM-iFEP) and Late Qualification for the Reading Assessment (N = 17,385), Controlling for Selected Child-Level and School-Level Demographics*

Variable	Symbol	Model 4: Child controls	Model 5: Child & school controls	Model 6: Final model	Model 6b: Accounting for PSUs with robust SEs
Fixed effects					
Final status					
Intercept	$\gamma_{00}$	142.484***	141.490***	141.379***	141.83***
LM-iLEP	$\gamma_{02}$	-2.493*	-1.282	-2.470 <sup>†</sup>	-2.968**
LM-iFEP	$\gamma_{03}$	2.415*	3.041**	2.484*	2.811**
iLEP—QUALTIME2	$\gamma_{04}$	-3.220	-3.261	-3.213	-2.821
iLEP—QUALTIME3	$\gamma_{05}$	-7.818***	-7.507***	-7.641***	-7.396***
iLEP—QUALTIME4	$\gamma_{06}$	-15.302***	-14.955***	-15.221***	-15.489***
FRLUNCH	$\gamma_{07}$		-0.080**	-0.092***	-0.091***
LM-iLEP × FRLUNCH	$\gamma_{08}$			0.072*	0.072*
LM-iFEP × FRLUNCH	$\gamma_{09}$			0.072*	0.080*
Final rate of change					
Intercept	$\gamma_{10}$	0.455***	0.432***	0.435***	0.464***
LM-iLEP	$\gamma_{12}$	0.095 <sup>†</sup>	0.070	0.166**	0.151**
LM-iFEP	$\gamma_{13}$	-0.012	-0.028	-0.047	-0.056 <sup>†</sup>
iLEP—QUALTIME2	$\gamma_{14}$	-0.223	-0.231 <sup>†</sup>	-0.236 <sup>†</sup>	-0.226*
iLEP—QUALTIME3	$\gamma_{15}$	-0.031	-0.041	-0.039	-0.027
iLEP—QUALTIME4	$\gamma_{16}$	0.048	0.031	0.044	-0.014
FRLUNCH	$\gamma_{17}$		0.001**	0.001**	0.001***
LM-iLEP × FRLUNCH	$\gamma_{18}$			-0.004**	-0.004***
LM-iFEP × FRLUNCH	$\gamma_{19}$			0.0006	0.001
Acceleration					
Intercept	$\gamma_{20}$	-0.021***	-0.021***	-0.021***	-0.021***
LM-iLEP	$\gamma_{22}$	0.002*	0.001	0.002*	0.002*
LM-iFEP	$\gamma_{23}$	-0.001 <sup>†</sup>	-0.002*	-0.002*	-0.002**
iLEP—QUALTIME2	$\gamma_{24}$	-0.006*	-0.006**	-0.006**	-0.006***
iLEP—QUALTIME3	$\gamma_{25}$	0.001	0.001	0.001	0.001
iLEP—QUALTIME4	$\gamma_{26}$	0.004 <sup>†</sup>	0.003	0.003 <sup>†</sup>	0.001
FRLUNCH	$\gamma_{27}$		0.00003***	0.00004***	0.00004***
LM-iLEP × FRLUNCH	$\gamma_{28}$			-0.00006**	-0.00007**
LM-iFEP × FRLUNCH	$\gamma_{29}$			-0.00002	-0.00001
Variance components					
Level 1					
Within child	$\sigma_e^2$	97.188***	96.825***	96.790***	134.676***
Level 2					
In intercept	$\sigma_0^2$	457.467***	450.640***	449.924***	216.701***
In rate of change	$\sigma_1^2$	0.068***	0.067***	0.066***	—
Covariance	$\sigma_{12}$	4.039***	3.964***	3.958***	—
Goodness of fit					
Deviance		517,445.54	516,932.98	516,881.32	—
Akaike information criterion		517,535.50	517,067.00	517,051.30	—
Bayesian information criterion		517,942.80	517,673.40	517,820.60	—
Likelihood ratio tests					
H <sub>0</sub>					
$\gamma_{08} = \gamma_{09} = \gamma_{18} = \gamma_{19} = \gamma_{28} = \gamma_{29} = 0$				51.66***	

*Note.* Each model also included centered final age as a predictor of final status, final rate of change, and acceleration. Child controls not shown included child-level socioeconomic status and ethnicity. School controls not shown included concentration of students of color and type of school (public, private—Catholic, private—religious, and private—other). Free and reduced lunch was centered at the sample mean. Therefore, the main effects of LM-iLEP and LM-iFEP in Models 6 and 6a can be interpreted as the effects of these variables in a school with an average concentration of poverty. LM = language minority; LM-iLEP = LM learners entered kindergarten with initially limited English proficiency; LM-iFEP = LM learners who entered kindergarten with initially full English proficiency; PSU = primary sampling units; QUALTIME2 = qualified at Time 2 to take the reading test; QUALTIME3 = qualified at Time 3 to take the reading test; QUALTIME4 = qualified at Time 4 to take the reading test; FRLUNCH = free or reduced-price lunch.

<sup>†</sup>  $p < .10$ . \*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

similar demographic backgrounds, the gap in reading achievement between students who were LM-iLEP and native English speakers was smaller among students in high-poverty schools than was the corresponding gap among students in low-poverty schools. Students who were LM-iFEP and attended high-poverty schools had fitted growth trajectories that were significantly higher in elevation than those of native English speakers in such schools, whereas these learners who attended low-poverty schools had fitted growth

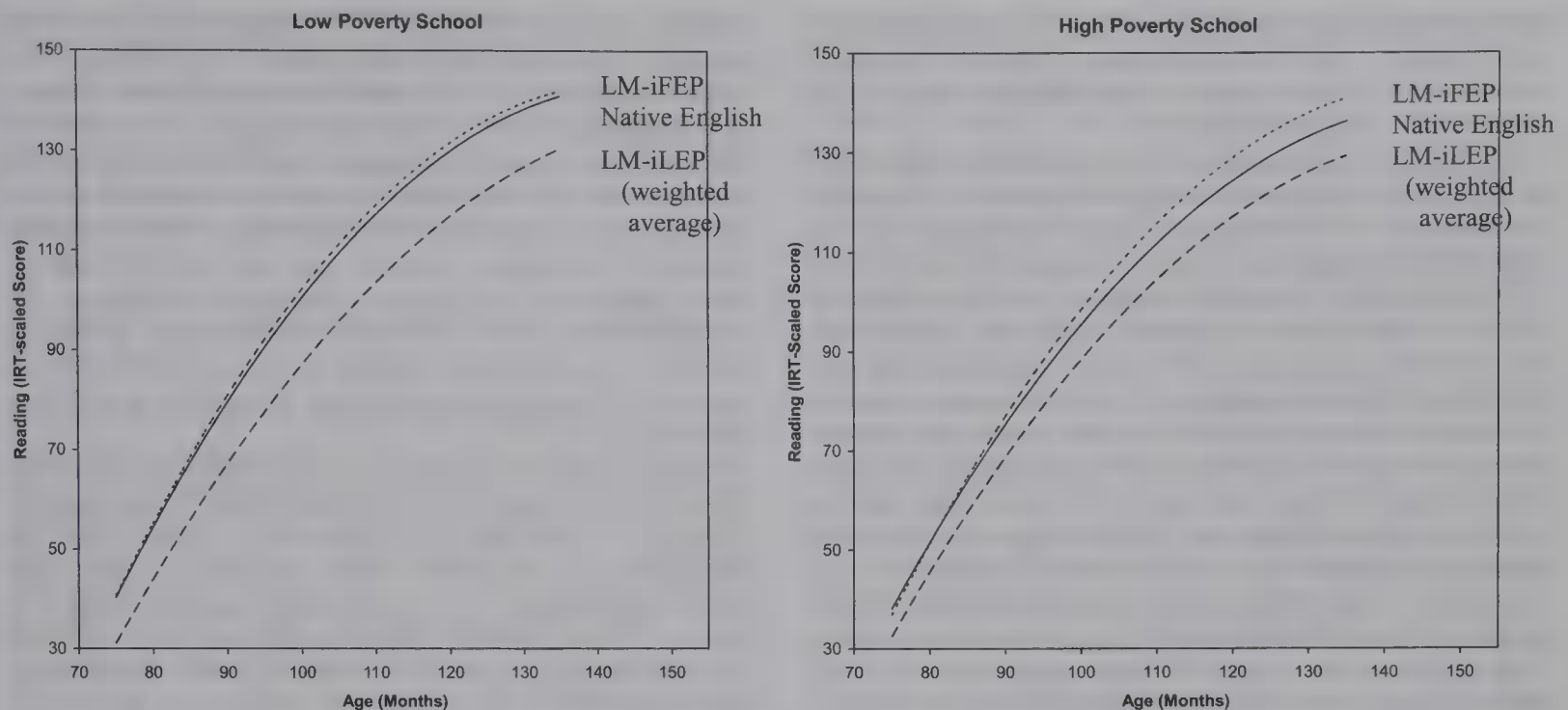


Figure 2. Fitted growth trajectories in reading for native English speakers, language minority learners who entered kindergarten with initially full English proficiency (LM-iFEP), language minority learners who entered kindergarten with initially limited English proficiency (LM-iLEP), who attend high-poverty and low-poverty schools, fitted for Latino children of average socioeconomic status attending public schools of average concentration of students of color ( $N = 17,385$ ). IRT = item response theory.

trajectories that were essentially equivalent to those of their native English speakers after controlling for individual demographics.

The two panels of Figure 2 display fitted growth trajectories in reading for students who attended low-poverty schools (in which 25% of students receive free or reduced-price lunch) and high-poverty schools (in which 75% of students receive free or reduced-price lunch). In both panels, the fitted trajectory of LM learners who were iLEP was a composite of the fitted trajectories of

students who were LM-iLEP, weighted by the proportion that qualified for the reading test in each round. Note that the fitted trajectories, and the effect sizes that follow, were for Latino students of average SES who attended public schools with an average concentration of students of color; given the interactions between age and these control variables in the model, these would be slightly different for a comparison involving prototypical cases with different values for the control variables. Table 6 provides the

Table 6

*Predicted Effect Sizes in Reading Achievement for Language Minority (LM) Learners Who Entered Kindergarten Proficient in English (iFEP) and LM Learners Who Entered Kindergarten With Limited English (iLEP) as Compared With Native English Speakers in High- and Low-Poverty Schools ( $N = 17,385$ )*

English language proficiency group	Effect size (Cohen's $d$ )			
	Low-poverty school		High-poverty school	
LM-iFEP				
Spring, Kindergarten	0.02		-0.09	
Spring, 3rd grade	0.07		0.16	
Spring, 5th grade	0.05		0.20	
LM-iLEP (weighted average)				
Spring, Kindergarten	-0.67		-0.41	
Spring, 3rd grade	-0.54		-0.29	
Spring, 5th grade	-0.45		-0.30	
LM-iLEP (by late qualifying)	Low estimate	High estimate	Low estimate	High estimate
Spring, Kindergarten	-0.42	—	-0.17	—
Spring, 3rd grade	-0.31	-0.88	-0.05	-0.63
Spring, 5th grade	-0.17	-0.83	-0.01	-0.67

Note. Each effect size was based on a fitted difference in the elevation of growth trajectories, fitted for Latino students of average socio-economic status who attended public schools with an average concentration of students of color. LM = language minority; LM-iLEP = LM learners entered kindergarten with initially limited English proficiency; LM-iFEP = LM learners who entered kindergarten with initially full English proficiency.



estimated differences, expressed as effect sizes, between native English speakers, LM learners who were iLEP, and LM learners who were iFEP in the spring of kindergarten, the spring of 3rd grade, and the spring of 5th grade.

Comparing the two panels of Figure 2 reveals that the differences between the English reading development of LM learners and that of native English speakers largely depended on the concentration of poverty in the schools they attended, as well as on the LM learners' initial proficiency in English. On one hand, among students who consistently attended low-poverty schools, the growth trajectories in reading of LM learners who entered kindergarten with limited English started lower and diverged more substantially from those of native English speakers with similar demographics, yielding moderate to large differences in fifth-grade achievement (with a weighted average effect size of .45). On the other hand, for students who attended high-poverty schools, the growth trajectories in reading for all students were lower in elevation (compared with those in low-poverty schools), but the differences between native English speakers and LM learners who were iLEP were substantially reduced (with an average effect size of .30 in fifth grade). For LM learners entering school proficient in English, the picture was quite different; in low-poverty schools, their trajectories had largely the same shape and elevation as those of native English speakers, but in high-poverty schools, their growth trajectories began at similar levels but diverged positively, yielding levels of reading achievement by fifth grade that were approximately .20 standard deviation higher.

Another way of interpreting this interaction is to say that differences in the growth trajectories in reading associated with attending schools of varying concentrations of poverty differed by students' English language proficiency group. Table 7 presents effect sizes for the predicted difference in reading achievement between students who attended high- and low-poverty schools over time by English language proficiency groups. Whereas stable and noteworthy differences existed between native English speakers who attended low- and high-poverty schools (between .16 standard deviation in kindergarten and .21 standard deviation in third grade), much smaller differences existed for students who were LM-iLEP and attended low- and high-poverty schools (rang-

ing from a positive difference of .10 standard deviation favoring students who attended high-poverty schools in kindergarten to a negative difference of .04 standard deviation favoring students who attended low-poverty schools by fifth grade). The pattern for LM learners who entered kindergarten proficient in English was even more striking. The association between school poverty and reading achievement diminished substantially over time for these learners. In kindergarten, students who were LM-iFEP and attended high-poverty schools had reading scores that were .27 standard deviation lower than those in low-poverty schools (as might have been expected), but this difference narrowed to .12 standard deviation by third grade and to .04 standard deviation by fifth grade.

Model 6b in Table 5 displays a version of the final model that accounted for the complex survey design while incorporating a simplified error structure. A comparison of Models 6 and 6b indicates that the magnitudes of the parameter estimates were largely similar whether or not the complex sampling design was taken into account. However, the two models did produce different results regarding the statistical significance of some small effects; most notably, Model 6b, in contrast to Model 6, indicated that students who were iLEP but qualified for the reading test by the end of kindergarten were significantly lower in fifth-grade status when compared with their native English-speaking counterparts of the same ethnicity and SES who attended schools of average SES. Nonetheless, these effect sizes were very similar in both specifications (−.11 and −.12), are considered small by the conventions of Cohen (1992), and have little practical importance, especially when compared with the much more substantial effects for students who were iLEP and qualified for testing later. Thus, the central findings were largely robust across these two methods of estimating effects.

Discussion

By examining language proficiency and English reading growth, this study provides several insights into our understanding of the reading development of LM learners. The first key finding is that the differences in levels of English reading across time between LM learners and native English speakers looked dramatically different when LM learners were disaggregated by initial English proficiency. LM learners who entered kindergarten proficient in oral English had substantial success in developing reading skills, whereas LM learners who entered kindergarten with LEP struggled throughout the elementary school years. The second key finding is that the answer to whether LM learners who entered kindergarten with LEP narrowed the gap over time depended on the group with which one compared their English reading achievement. Controlling for demographic risk factors, most notably the large effect of SES, the magnitude of the effect of initial English proficiency was reduced from large to moderate. Moreover, although the average growth trajectory in reading of students who were LM-iLEP diverged from the national average for native English speakers between kindergarten and fifth grade, the average trajectory of students who were LM-iLEP converged with that of students who shared demographic risk factors (i.e., were of the same ethnicity and SES). The third key finding is that LM learner status moderated the negative effect of attending a high-poverty school. The differences in reading

Table 7  
*Predicted Effect Sizes in Reading Achievement Between Students Attending Schools With High and Low Concentrations of Poverty by English Language Proficiency Group (N = 17,385)*

Grade	Effect sizes (Cohen's <i>d</i> )		
	Native English speakers	iFEP students	iLEP students (weighted average)
Spring, Kindergarten	−0.16	−0.27	0.10
Spring, 3rd grade	−0.21	−0.12	0.04
Spring, 5th grade	−0.20	−0.04	−0.04

*Note.* Positive differences favor low-poverty schools (in which 25% of students receive free or reduced lunch) in comparison with high-poverty schools (in which 75% of students receive free or reduced lunch). Effect sizes are fitted for Latino students of average socioeconomic status attending public schools of average concentration of students of color. iLEP = initially limited English proficiency; iFEP initially full English proficiency.



trajectories between students who attended high- and low-poverty schools were smaller for both groups of LM learners than for native English speakers. Moreover, the trajectories of LM learners in high-poverty schools tended to converge over time on those of LM learners in low-poverty schools (unlike the growth trajectories of native English speakers who attended schools of differing concentrations of poverty).

### *Reading and Language Development in English: Who's Falling Behind?*

The results of this study suggest that LM learners in the United States who enter kindergarten orally proficient in English reach levels of English reading achievement that are equivalent to or higher than those reached by native English speakers. By fifth grade their English reading achievement was not significantly different from the national average and was approximately .16 standard deviation higher than demographically similar native English speakers in high-poverty schools. On the other hand, LM learners who entered kindergarten with limited English proficiency had large, persistent deficiencies in English reading achievement. Although the estimates of these gaps depended substantially on whether demographic risk factors were taken into account and on the stage at which students gained enough English proficiency to take the reading test, none of the estimates of these gaps were small. Even the students who acquired English most rapidly, in the course of a year of kindergarten, continued to lag behind the national average for native English speakers by more than .33 standard deviation in third and fifth grade.

The strong performance of LM learners in this study who entered kindergarten fluent in English both reinforces the research base on bilingualism and refocuses attention on the issue of English proficiency. These findings suggest that exposure to a language other than English before kindergarten does not, by itself, place students on a differential trajectory for English reading achievement. This concurs with researchers who have often argued that bilingualism is not inherently detrimental to reading development (e.g., August & Hakuta, 1997; Snow, 1992); rather, it suggests that limited English proficiency and the constrained access to the school curriculum it entails are the characteristics that put students at elevated risk for English reading difficulties. It also suggests U.S. schools may not be failing all bilingual students, but that the most underserved population consists of those who enter school with limited English skills.

Additionally, in describing the English reading growth of LM learners who enter kindergarten with limited English, these findings support the differential skills hypothesis and related theoretical models (e.g., RAND Reading Study Group, 2002). For these learners, moderate deficiencies in reading at the end of kindergarten grow into large deficiencies as text demands increase, likely in part because of their persistently low vocabulary levels. Stanovich (1986) has hypothesized that a reciprocal relationship between reading ability and vocabulary knowledge underlies individual differences in reading that increase over time, especially during the late elementary grades as students transition from learning to read, a stage in which they rely largely on basic vocabulary acquired through conversation, to reading to learn, a stage at which readers are expected to acquire sophisticated vocabulary and content

knowledge through reading (Chall, 1983; RAND Reading Study Group, 2002).

For LM learners who enter kindergarten with limited vocabulary knowledge in English, this developmental process can be a downward spiral that accelerates in later years, as students not only lack the vocabulary to comprehend and analyze texts, but then fail to gain the vocabulary and knowledge from texts that are essential to later success with English reading. Even the group of students in the present analysis who nominally acquired oral English proficiency by the end of first grade (as indicated by passing the Pre-K LAS assessment) continued to have slow rates of growth in English reading across the elementary school years. This supports the assertion of second-language researchers that acquiring basic interpersonal English is a simpler task than acquiring the academic language and vocabulary necessary to comprehend academic texts in English (e.g., Collier, 1989; Cummins, 1979; Scarcella, 2004). The difficulty of the latter task is compounded by the limited instructional time that is typically devoted to vocabulary and language development (e.g., Roser & Juel, 1982; Scott, Jamieson-Noel, & Asslin, 2003; Watts, 1995).

A limitation of the present study is that it used only a single global measure of English reading and thus could not support the examination of differential development in componential reading skills. Future large-scale longitudinal research of this kind should consider including standardized measures of multiple constructs related to reading, such as word reading, vocabulary, and comprehension processes. Such research should also consider including assessments of native language skills and collecting data on students' opportunities to develop native language skills, although such assessments should supplement rather than replace assessments in English to provide a complete picture of students' achievement.

### *Incorporating Demographic Factors: Catching Up to Whom?*

The second key finding is that the growth trajectories in reading of LM learners who entered kindergarten with limited English proficiency diverged from the national average for native English speakers but converged with those of native English speakers with whom they shared demographic risk factors. When ethnicity and SES were not taken into account, differences in English reading between native English speakers and students who were LM-ILEP grew between kindergarten and fifth grade (from approximately .60 standard deviation to .80 standard deviation). However, when ethnicity, SES, and school demographics were taken into account, these differences narrowed between kindergarten and fifth grade (from approximately .60 standard deviation to less than .40 standard deviation in schools of average concentration of poverty).

Given that the reading achievement of native English speakers with multiple demographic risk factors who attended high-poverty schools was well below the national average, it is unfortunate that LM learners who entered kindergarten with limited English did not surpass this low standard. Nonetheless, this finding echoes the suggestion of Lesaux et al. (2007) and others that the confounding of SES and LM learner status in many settings may obscure which part of the achievement gap is attributable to LM learner status. It also indicates that the levels of English reading achievement among many LM learners in the upper elementary grades may be fairly similar to those of their



classmates. Future research should investigate the degree to which these two groups' academic needs, instructional profiles, and responsiveness to intervention also overlap.

### *The Moderating Effect of LM Learner Status on School Poverty*

The third key finding is that the effects of school poverty differed as a function of LM learner status, such that differences between growth trajectories in reading associated with school poverty were greater for native English speakers than for LM learners. These findings suggest that the negative effects of concentrated poverty are actually less severe for LM learners than for native English speakers. This somewhat surprising interaction effect can be interpreted in several ways.

One interpretation of this moderating effect is that LM learners have resilience to the effect of school poverty because of resources associated with their LM learner status, such as cognitive advantages provided by bilingualism or unobserved family resources. This interpretation may explain the success of LM learners who enter school proficient in English; although the subgroup of these learners who entered high-poverty schools was lower than their counterparts in low-poverty schools by .25 standard deviation in the spring of kindergarten, they reached equivalent levels by the spring of fifth grade. As displayed in Figure 2, students who were LM-iFEP reached high reading levels by fifth grade whether they attended high- or low-poverty schools. This success represented levels of achievement that were equivalent to those of their native-English speaking classmates in low-poverty schools and higher levels than those of such classmates in high-poverty schools.

Future research is certainly needed to investigate the sources of this resilience. Two explanations are worth considering. First, given that this group of learners entered kindergarten proficient in English and with at least some proficiency in another language, we might suspect that they have cognitive or linguistic advantages because of being raised in a bilingual home. This explanation draws empirical support largely from the work of Bialystok (1988, 2005), who has argued that bilingual children have advantages in specific reading-related skills, particularly in tasks involving meta-cognitive and meta-linguistic awareness. One aspect of meta-linguistic awareness that is particularly important in the process of learning to read for both native English speakers and LM learners is phonological awareness (for reviews, see Adams, 1990; August & Shanahan, 2006; National Institute of Child Health & Human Development, 2000). However, there is mixed evidence as to whether bilingualism facilitates development of phonological awareness. Bialystok, Majumder, and Martin (2003) found advantages for Spanish-English bilinguals on one phonological awareness task but failed to replicate this finding when using different tasks or among speakers of other languages, whereas other studies have found average to below average levels of phonological awareness among LM learners (e.g., Lesaux & Siegel, 2003; Manis et al., 2004).

A second explanation for the high reading levels attained by students who were iFEP may be the presence of family resources associated with recent immigration that are not adequately captured with a single measure of SES. For instance, the National Research Council (1998) concluded that children in families who have recently immigrated have overall better physical and psycho-

logical health and engage in fewer risky behaviors than their counterparts in second- and third-generation immigrant families. Similarly, in their study of children of immigrants, Suarez-Orozco and Suarez-Orozco (2001) found that an emphasis on parents' current occupation and salary obscured the bimodal distribution of other resources among immigrant families, including parental educational levels; whereas some parents immigrated to the United States because of their own limited educational and occupational opportunities in their home countries, others had high levels of education that allowed them to navigate the process of immigration. Such resources and related family behaviors may serve as protective factors against the negative effects of attending high-poverty schools for some LM learners and thus explain differences within the population of LM learners. In examining heterogeneity among Spanish-speaking LM learners, Reese et al. (2000) found that family literacy practices and grandparents' education predicted seventh-grade reading outcomes. Although this evidence suggests that many immigrant families may have resources that explain the resilience of some LM learners to school poverty, more research is needed to examine the links between these resources and English reading development.

Similarly, the negative effects of school poverty may be less severe for LM learners who enter kindergarten with limited English proficiency. However, the consistently low performance of these learners in both high- and low-poverty schools might best be characterized as a low baseline of depressed development. Unlike those students who were LM-iFEP and entered low-poverty schools with average English reading achievement, students who were LM-iLEP entered school with low levels of English reading, regardless of whether they entered high- or low-poverty schools. Over time, they maintained equivalently low levels despite differences in their schooling contexts, which is more suggestive of the difficulties of these learners in low-poverty schools across the elementary school years than the resilience of their counterparts in high-poverty schools.

It is possible that this subset of LM learners does not benefit from the same linguistic and family resources as do LM learners who are bilingual at school entry. However, it may also be that low-poverty schools underserve these learners compared with LM learners with higher initial English language and literacy levels. Specific instructional programs to develop English proficiency may be scarce in schools that lack funding from Title I or Title III. Teachers in such schools may be less prepared to serve the needs of students with limited English proficiency or may have low expectations for these learners. Alternatively, given the critical role of language in reading development, it is possible that students who enter schools with limited English proficiency cannot attain a threshold of English proficiency at which they could benefit from the resources that a low-poverty school has to offer (at least in the absence of early and extensive intervention). Future research should investigate the instructional and contextual factors at play in LM learners' English reading development in low-poverty schools, especially given the limited research on these learners outside of urban centers and the growing numbers of LM learners in schools of all demographic constitutions (Capps et al., 2005).

In summary, this study suggests that students who enter school with limited proficiency in English are at great risk for reading difficulties and have pressing instructional needs that are currently not addressed in elementary schools in the United States. Further-



more, the findings illustrate the great risk also experienced by other students of color who come from low-SES backgrounds and suggest that the academic needs of these two groups of learners may be quite similar, particularly in high-poverty schools.

## References

- Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.
- Anderson, R. C., & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. In B. Hutson (Ed.), *Advances in reading/language research* (Vol. 2, pp. 231–256). Greenwich, CT: JAI Press.
- August, D. E., & Hakuta, K. E. (1997). *Improving schooling for language-minority children: A research agenda*. Washington, DC: National Academy Press.
- August, D. E., & Shanahan, T. (2006). *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth*. Mahwah, NJ: Erlbaum.
- Bialystok, E. (1988). Levels of bilingualism and levels of linguistic awareness. *Developmental Psychology*, 24, 560–567.
- Bialystok, E. (2005). Consequences of bilingualism for cognitive development. In J. Kroll (Ed.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 417–432). New York: Oxford University Press.
- Bialystok, E., Majumder, S., & Martin, M. (2003). Developing phonological awareness: Is there a bilingual advantage? *Applied Linguistics*, 24, 27–44.
- Capps, R., Fix, M., Murray, J., Ost, J. Passel, J. S., & Herwanto, S. (2005). *The new demography of America's schools: Immigration and the No Child Left Behind Act* (Research report). Washington, DC: Urban Institute.
- Chall, J. (1983). *Stages of reading development*. New York: McGraw Hill.
- Chatterji, M. (2006). Reading achievement gaps, correlates, and moderators of early reading achievement: Evidence from the Early Childhood Longitudinal Study (ECLS) kindergarten to first grade sample. *Journal of Educational Psychology*, 98, 489–507.
- Chiappe, P., Siegel, L. S., & Wade-Woolley, L. (2002). Linguistic diversity and the development of reading skills: A longitudinal study. *Scientific Studies of Reading*, 6, 369–400.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155–159.
- Cohen, J. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Mahwah, NJ: Erlbaum.
- Collier, V. P. (1989). How long? A synthesis on academic achievement in a second language. *TESOL Quarterly*, 23, 509–531.
- Compton, D. L. (2000). Modeling the growth of decoding skills in first-grade children. *Scientific Studies of Reading*, 4, 219–259.
- Cosentino de Cohen, C., Deterding, N., & Chu Clewell, B. (2005). *Who's left behind? Immigrant children in high and low LEP schools* (Policy Report). Washington, DC: Urban Institute.
- Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimal age question and some other matters. *Working Papers on Bilingualism*, 19, 121–129.
- Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology*, 33, 934–945.
- Duncan, S. E., & DeAvila, E. A. (1998). *Pre-Language Assessment Scale 2000*. Monterey, CA: CTB/McGraw-Hill.
- Foorman, B. R., Francis, D. J., Novy, D. M., & Liberman, D. (1991). How letter-sound instruction mediates progress in first-grade reading and spelling. *Journal of Educational Psychology*, 83, 456–469.
- Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag verse deficit models of reading disability: A longitudinal, growth curves analysis. *Journal of Educational Psychology*, 88, 3–17.
- Gándara, P., Rumberger, R., Maxwell-Jolly, J., & Callahan, R. (2003). English learners in California schools: Unequal resources, unequal outcomes. *Education Policy Analysis Archives*, 11(36). Retrieved February 26, 2007, from <http://epaa.asu.edu/epaa/v11n36/>
- Gerber, M., Jimenez, T., Leafstedt, J., Villaruz, J., Richards, C., & English, J. (2004). English reading effects of small-group intensive intervention in Spanish for K-1 English learners. *Learning Disabilities Research & Practice*, 19, 239–251.
- Hakuta, K., Butler, Y. G., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* (Policy report). Santa Barbara: University of California Language Minority Research Institute.
- Lesaux, N. K. (with Koda, K., Siegel, L. S., & Shanahan, T.). (2006). Development of literacy. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth* (pp. 75–122). Mahwah, NJ: Erlbaum.
- Lesaux, N. K., Rupp, A., & Siegel, L. S. (2007). Growth in reading skills of children from diverse linguistic backgrounds: Findings from a 5-year longitudinal study. *Journal of Educational Psychology*, 99, 821–834.
- Lesaux, N. K., & Siegel, L. S. (2003). The development of reading in children who speak English as a second language. *Developmental Psychology*, 39, 1005–1019.
- Manis, F., Lindsey, K., & Bailey, A. (2004). Development of reading in grades K-2 in Spanish-speaking English-language learners. *Learning Disabilities Research & Practice*, 19, 214–224.
- National Center for Educational Statistics. (2007). *Nation's report card: Reading, 2007*. Washington, DC: Author.
- National Institute of Child Health & Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00–4769). Washington, DC: U.S. Government Printing Office.
- National Research Council. (1998). *From generation to generation: The health and well-being of children in immigrant families*. Washington, DC: National Academy Press.
- O'Malley, K. J., Francis, D. J., Foorman, B. R., Fletcher, J. M., & Swank, P. R. (2002). Growth in precursor and reading-related skills: Do low-achieving and IQ-discrepant readers develop differently? *Learning Disabilities Research & Practice*, 17, 19–34.
- Perfetti, C. A. (1988). Verbal efficiency in reading ability. In M. Daneman, G. E. MacKinnon, & T. G. Waller (Eds.), *Reading research: Advances in theory and practice* (pp. 109–143). New York: Academic Press.
- Pollack, J. M., Atkins-Burnett, S., Najarian, M., & Rock, D. A. (2005). *Early Childhood Longitudinal Study, kindergarten class of 1998–99 (ECLS-K) psychometric report for the fifth grade* (NCES 2006036). Washington, DC: National Center for Education Statistics.
- RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND Corporation.
- Reese, L., Garnier, H., Gallimore, R., & Goldenberg, C. (2000). Longitudinal analysis of the antecedents of emergent Spanish literacy and middle-school English reading achievement of Spanish-speaking students. *American Educational Research Journal*, 37, 633–662.
- Roser, N., & Juel, C. (1982). Effects of vocabulary instruction on reading comprehension. In J. A. Niles & L. A. Harris (Eds.), *Yearbook of the National Reading Conference. Vol. 31: New inquiries in reading research and instruction* (pp. 110–118). Rochester, NY: National Reading Conference.
- Scarcella, R. (2004). *Academic English: A conceptual framework* (University of California Linguistic Minority Research Institute Technical Report 2003–1). Santa Barbara: University of California Linguistic Minority Research Institute. Retrieved March 1, 2007, from [http://lmri.ucsb.edu/publications/03\\_scarcella.pdf](http://lmri.ucsb.edu/publications/03_scarcella.pdf)
- Scott, J. A., Jamieson-Noel, D., & Asselin, M. (2003). Vocabulary instruc-



- tion throughout the day in twenty-three Canadian upper-elementary classrooms. *Elementary School Journal*, 103, 269–283.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. New York: Oxford University Press.
- Snow, C. E. (1992). Perspectives on second-language development: Implications for bilingual education. *Educational Researcher*, 21, 16–19.
- Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.
- Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, 21, 360–407.
- Stuart, M. (2004). Getting ready for reading: A follow-up study of inner city second language learners at the end of Key Stage 1. *British Journal of Educational Psychology*, 74, 15–36.
- Suarez-Orozco, C., & Suarez-Orozco, M. (2001). *Children of immigration*. Cambridge, MA: Harvard University Press.
- Swanson, H. L., & Jerman, O. (2007). The influence of working memory on reading growth in subgroups of children with reading disabilities. *Journal of Experimental Child Psychology*, 96, 249–283.
- Swanson, H. L., Saez, L., & Gerber, M. (2006). Growth in literacy and cognition in bilingual children at risk or not at risk for reading disabilities. *Journal of Educational Psychology*, 98, 247–264.
- Tourangeau, K., Lê, T., & Nord, C. (2005). *Early Childhood Longitudinal Study, kindergarten class of 1998–99 (ECLS-K) fifth-grade methodology report* (NCES 2006037). Washington, DC: National Center for Education Statistics.
- Vellutino, F. R., Scanlon, D. M., Small, S. G., & Tanzman, M. S. (1991). The linguistic basis of reading ability: Converting written to oral language. *Text*, 11, 99–133.
- Verhoeven, L. T. (1990). Acquisition of reading in a second language. *Reading Research Quarterly*, 25, 90–114.
- Verhoeven, L. T. (2000). Components in early second language reading and spelling. *Scientific Studies of Reading*, 4, 313–330.
- Watts, S. M. (1995). Vocabulary instruction during reading lessons in six classrooms. *Journal of Reading Behavior*, 27, 399–424.
- Zehler, A. M., Fleischman, H. L., Hopstock, P. J., Stephenson, T. G., Pendzick, M. L., & Sapru, S. (2003). *Descriptive study of services to LEP students and LEP students with disabilities. Vol. 1: Research Report*. Arlington, VA: Development Associates.

Received July 10, 2007

Revision received April 30, 2008

Accepted May 13, 2008 ■

### New Editors Appointed, 2010–2015

The Publications and Communications Board of the American Psychological Association announces the appointment of 4 new editors for 6-year terms beginning in 2010. As of January 1, 2009, manuscripts should be directed as follows:

- *Psychological Assessment* (<http://www.apa.org/journals/pas>), **Cecil R. Reynolds, PhD**, Department of Educational Psychology, Texas A&M University, 704 Harrington Education Center, College Station, TX 77843.
- *Journal of Family Psychology* (<http://www.apa.org/journals/fam>), **Nadine Kaslow, PhD**, Department of Psychiatry and Behavioral Sciences, Grady Health System, 80 Jesse Hill Jr. Drive, SE, Atlanta, GA 30303.
- *Journal of Experimental Psychology: Animal Behavior Processes* (<http://www.apa.org/journals/xan>), **Anthony Dickinson, PhD**, Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, United Kingdom
- *Journal of Personality and Social Psychology: Personality Processes and Individual Differences* (<http://www.apa.org/journals/psp>), **Laura A. King, PhD**, Department of Psychological Sciences, University of Missouri, McAlester Hall, Columbia, MO 65211.

**Electronic manuscript submission:** As of January 1, 2009, manuscripts should be submitted electronically via the journal's Manuscript Submission Portal (see the website listed above with each journal title).

Manuscript submission patterns make the precise date of completion of the 2009 volumes uncertain. Current editors, Milton E. Strauss, PhD, Anne E. Kazak, PhD, Nicholas Mackintosh, PhD, and Charles S. Carver, PhD, will receive and consider manuscripts through December 31, 2008. Should 2009 volumes be completed before that date, manuscripts will be redirected to the new editors for consideration in 2010 volumes.

# From Reading to Spelling and Spelling to Reading: Transfer Goes Both Ways

Nicole J. Conrad  
Saint Mary's University

This study compares the effects of practice spelling and reading specific words on the orthographic representations in memory involved in reading both practiced words and new, unfamiliar words. Typically developing readers in Grade 2 (mean age = 7 years, 7 months) participated in a training study examining whether transfer can occur between reading and spelling following a series of reading and spelling practice sessions. Practice consisted of either repeated reading or repeated spelling of words with shared orthographic rime patterns. A series of mixed analyses of variance was used to examine generalization within skill and transfer across skill. Following practice, word-specific transfer across skill was found. Specifically, children were better able to spell words they had practiced reading and to read words they had practiced spelling. In addition, generalization to new words with practiced rime units was found both within a skill and across skills. However, transfer from spelling to reading was greater than transfer from reading to spelling. Results indicate that the orthographic representations established through practice can be used for both reading and spelling. Subsequently, reading and spelling curricula should be coordinated to benefit children maximally.

*Keywords:* reading, spelling, orthographic knowledge, intervention

Fluent reading rests on word recognition, among other skills. If difficulty is encountered recognizing individual words, few, if any, higher order skills such as comprehension can operate (Laberge & Samuels, 1974; Snowling, 2000). Many models of reading propose that fluent reading of words results from the rapid recognition of letter sequences as a single unit within a word (Adams, 1990; Ehri, 1997). This rapid recognition depends on previously established memory representations of letter patterns that occur frequently within a language (i.e., orthographic representations). Memory representations of these letter sequences that are accurate and detailed contain information about the graphemes (i.e., letters), phonemes (i.e., sounds), and their mappings (Ehri, 1997) and are thought to be established through practice with words. In the present study, I compare two types of practice through which children may establish the orthographic representations necessary for fluent reading. Specifically, I examine whether the orthographic representations established through practicing spelling words may facilitate reading words.

Models of reading suggest that the orthographic memory representations necessary for fluent reading are established primarily through the repeated phonological recoding of words (Ehri, 1997,

2005; Share, 1995). Repeated practice mapping phonemes onto graphemes allows children to set up the representations necessary to support fluent word reading (Cunningham, Perry, Stanovich, & Share, 2002; Share, 2004). With continued practice, these memory representations become increasingly accurate, reflecting the constituent letter sequences of words and their phonological code (Perfetti & Hart, 2001). The idea that practice is necessary for the improvement of reading skills is not new. For instance, Nathan and Stanovich (1991) declared more than a decade ago that children develop knowledge of orthographic patterns through practice with words, a conclusion supported by the finding that children who read a lot become better readers, whereas children who read little fall further and further behind their peers (Stanovich, 1986).

It has been assumed in the research literature that the best practice takes the form of repeated reading (for a review, see Meyer & Felton, 1999). Practice at reading words enables children to recognize common letter patterns, to map sounds onto letters and letter patterns, and to develop fully specified orthographic representations. In fact, according to the self-teaching hypothesis (Share, 1995, 2004), translating a printed word into its spoken equivalent is the only means through which fully specified orthographic representations can be established. Several studies have supported the important role of repeated reading for improving word recognition skills in terms of both accuracy and speed (e.g., Berends & Reitsma, 2006; Levy, 2001; Thaler, Ebner, Wimmer, & Landerl, 2004). These studies focused on providing practice with a set of specific orthographic patterns (e.g., *-uck*) through repeated reading of words containing these patterns. Their goal was to use such practice to set up orthographic representations that might facilitate reading new, nonpracticed words that contained the same orthographic patterns. For example, Levy (2001) reported a study in which children with poor reading skills in Grade 2 were pre-

---

This research was supported by a grant provided by the Brandon University Research Committee. A special thank you to the principals, teachers, and children who participated in this research and to Sabrina Cutting for help with data collection. Portions of this article were presented at the Annual Meeting of the Society for the Scientific Study of Reading, Toronto, Canada, in June 2005.

Correspondence concerning this article should be addressed to Nicole J. Conrad, Department of Psychology, Saint Mary's University, 923 Robie Street, Halifax, Nova Scotia, Canada B3H 3C3. E-mail: nicole.conrad@smu.ca



sented with words that shared a rime unit (e.g., *-uck*) blocked together for repeated reading practice. Results indicated a benefit in reading accuracy for the actual words children had practiced reading, but little generalization to new words that contained the practiced orthographic patterns. Similar results using the same methodology have been found with Dutch children (Berends & Reitsma, 2006) and German children (Thaler et al., 2004) with poor reading skills. Together, the results of these training studies suggest that repeated reading is effective in setting up word-specific orthographic representations that facilitate reading those specific words. However, another measure of the success of a training program lies not only in its word-specific effects but also in how well children can generalize what they have learned through training to reading new words. As the studies reviewed here found little generalization to new material, there is a need to explore other methods of practice that might establish orthographic representations that can aid in reading new or unfamiliar words.

The current study explored whether spelling practice would contribute to the establishment of orthographic representations that facilitate reading. Both Ehri (1997, 2005) and Perfetti (1997) proposed that reading and spelling share mental representations, or knowledge sources, suggesting that what is learned through one skill should benefit the other skill (e.g., Ehri, 1997; Perfetti, 1997), albeit not to the same degree. Spelling is the more difficult task and requires a better quality representation (Perfetti, 1992). Thus, although reading can be supported by incomplete representations, spelling cannot. For example, individuals can read the word *occasion* even if their representation contains the letter pattern *oc?a-s?ion*, whereas this representation might lead to an incorrect spelling (Perfetti, 1997; Perfetti & Hart, 2001). Like reading, spelling provides practice mapping phonemes onto graphemes, but because it requires the production rather than the recognition of letter patterns, full attention to all letters in a word is necessary. In addition, the motor act involved in spelling may provide an additional cue that is incorporated into the memory representation (Hulme, 1981). This increased thoroughness of processing letters may produce representations that are more complete than those established through reading alone. Thus, a high-quality representation established through practice spelling a word should support reading that same word; however, the quality of the representation established from practice reading a word would not necessarily ensure that word could be spelled.

Few studies have examined word-specific transfer from reading to spelling and spelling to reading concurrently, preventing a direct comparison of the effectiveness of different types of practice. However, a number of studies have examined the benefit to spelling of practicing reading. For example, Share (2004) found that practice reading Hebrew words in a text facilitated spelling those same words for children in Grade 3 but not for children in Grade 1. In the English language, Ehri has conducted a number of studies in which practice reading nonwords (Ehri, 1980) and words (Ehri & Wilce, 1986) has enhanced the spelling of those nonwords and words for children in Grade 2. On the basis of these findings, Ehri (1997) concluded that reading practice sets up word-specific representations that can support spelling those words. However, the amount of transfer from reading to spelling was not 100%. Following practice reading nonwords, Ehri (1980) found that 69% of those nonwords were spelled correctly; following practice reading real words (Ehri & Wilce, 1986), only 31% of practiced words

were spelled correctly. Other studies have failed to find any effects on spelling following repeated reading practice. A series of studies reported by Bosman and van Orden (1997) have suggested that reading practice is not the most effective way to learn to spell. In these studies, more repetitions of reading words did not result in a corresponding improvement in spelling those words for a group of beginning Dutch readers. Because of these conflicting results, it is unclear how reading practice benefits spelling.

Our understanding of how practice spelling benefits reading is also incomplete. Teaching general alphabetic knowledge through spelling, such as mapping phonemes onto graphemes, has had some influence on reading skills (e.g., Berninger et al., 1998; Ehri & Wilce, 1987). However, when word-specific spelling practice is used (repeated spelling of a set of words), there appears to be little benefit to reading those words. For example, Roberts and Ehri (1983) tested children in Grade 2 who practiced spelling a list of pseudowords through two different techniques. Although the groups differed in spelling accuracy for those pseudowords following spelling practice, the groups did not differ in the number of pseudowords read correctly or in the speed of reading those pseudowords following spelling practice. Although the results might suggest that what is learned through spelling does not transfer to reading, such a conclusion is inconsistent with theories of reading and spelling (e.g., Ehri, 1997; Perfetti, 1997). In fact, Ehri (1997) suggested that further study is needed to understand fully whether word-specific information gained through spelling practice can benefit reading.

Orthographic representations established through practice with words can also support the reading of unfamiliar words, either through analogy from known words containing shared letter sequences (e.g., Ehri & Robbins, 1992; Goswami, 1986) or through the induction of sublexical relations (e.g., Fletcher-Flinn & Thompson, 2004; Thompson, Cottrell, & Fletcher-Flinn, 1996) that provide readers with general knowledge of the relations between orthographic and phonological components that are common to words the child has experienced. Whereas the above-mentioned studies examining generalization following repeated reading of words with poor readers have failed to find meaningful generalization to reading new words (e.g., Berends & Reitsma, 2006; Levy, 2001; Thaler et al., 2004), studies using different methodologies have indicated that generalization to reading unfamiliar words (e.g., Goswami, 1986) after practice reading and to spelling unfamiliar words (e.g., Bosse, Valdois, & Tainturier, 2003; Campbell, 1985) following practice spelling can occur. However, whether generalization to new words can occur across skills (i.e., from spelling to reading or reading to spelling) has not been directly examined in the literature.

### The Present Study

Presently, our understanding of the transfer between reading and spelling is incomplete. Theoretically, practice spelling should benefit reading more than practice reading should benefit spelling, yet studies to date have not allowed a direct comparison of the reciprocal benefits. The current study explores these issues in a group of typically developing readers to examine the nature of transfer between skills and the degree to which it can occur. Children received either repeated practice spelling or repeated practice reading a list of words containing families of words with shared



orthographic rime units. To examine both generalization within and across skill following practice, children were asked to read and spell a new list of words, some of which contained the practiced orthographic pattern. To compare word-specific transfer across skill, children also spelled the words they had practiced reading or read the words they had practiced spelling. Children were also matched on a number of skills related to reading and spelling, including vocabulary (e.g., Beck & McKeown, 1991), orthographic knowledge (e.g., Barker, Torgesen, & Wagner, 1992), and phonological skill (e.g., Bradley & Bryant, 1983), to reduce the likelihood that any benefits after practice were the result of pre-existing differences in these skills. This design enabled several questions related to the benefits of each type of practice to be addressed.

Because spelling a word requires a more complete representation than does reading a word, spelling practice should establish a more detailed and accurate orthographic representation than reading practice, which results in the following four predictions: First, generalization within skill was expected after both reading and spelling practice because typically developing readers in Grade 2 have the decoding skills necessary for making analogies and inducing the sublexical relations between graphemes and phonemes (Ehri & Robbins, 1992). Second, generalization to spelling unfamiliar words after spelling practice should be greater than to reading unfamiliar words after reading practice because of the more accurate and detailed orthographic representation established through spelling practice. Third, word-specific transfer from spelling to reading should be greater than from reading to spelling because accurately spelling a word should ensure that word can be read, whereas accurately reading a word does not necessarily ensure that word can be spelled. Fourth, transfer to unfamiliar words from spelling to reading should be greater than from reading to spelling because of the more accurate and detailed orthographic representation that readers have for reading and spelling new words.

An investigation into these issues can not only provide us with information regarding the nature of the representations that underlie reading and spelling, but perhaps more important, results can inform curricular design as to how to coordinate reading and spelling instruction to benefit students maximally.

## Method

### Participants

Forty-one children (22 boys and 19 girls) in Grade 2, with a mean age of 7 years, 7 months, participated. They were selected from 60 children screened in five schools in a rural area in western Canada. The average yearly income in these rural areas, according to 2000 census data, was \$23,449, below the national average of \$43,298 (Statistics Canada, 2002). All children spoke English as a first language. Each child was administered the Word Identification subtest of the Woodcock-Johnson Tests of Achievement—Third Edition (WJ-III; Woodcock, McGrew, & Mathers, 2001). Forty-four children scoring below a standard score of 115 and above a standard score of 85 (1 standard deviation from the mean) on the WJ-III were identified as average readers and included in the study. Of these 44 children, 1 was identified as having attention deficit/hyperactivity disorder and excluded from data analyses, and

2 were excluded because of noncompliance or incomplete data. Thus, the final sample consisted of 41 children, 20 children in the reading practice group (readers) and 21 children in the spelling practice group (spellers).

### Pretest Measures

**Reading skill.** Reading skill was measured using the Word Identification subtest of the WJ-III. This is a standardized test in which children read a list of words that become progressively more difficult (e.g., *see*, *bought*, and *investigate*). Testing is discontinued after six errors in a row. Reported corrected split-half reliability for 7-year-olds on this subtest is .97. Dependent measures used in data analyses were standard scores (age normed), with a mean of 100 and standard deviation of 15.

**Spelling skill.** Spelling skill was measured using the Spelling subtest of the WJ-III. This is a standardized test measuring written spelling. Words are orally presented to the child and become progressively more difficult (e.g., *hat* and *league*). Testing is discontinued after six errors in a row. Reported corrected split-half reliability for 7-year-olds on this subtest is .91. Dependent measures used in data analyses were standard scores (age normed) with a mean of 100 and a standard deviation of 15.

**Vocabulary.** Receptive vocabulary was assessed using the Picture Vocabulary subtest of the WJ-III. Children are asked to identify pictures of objects (e.g., fish, umbrella, and moccasins), and testing is discontinued after six errors in a row. Reported corrected split-half reliability for 7-year-olds on this subtest is .71. Dependent measures used in data analyses were standard scores (age normed) with a mean of 100 and a standard deviation of 15.

**Phonological awareness.** The Test of Auditory Analysis Skills (Rosner, 1979) was administered to measure phonological awareness. This test is a 13-item sound-deletion task in which children are to repeat a word spoken by the experimenter after deleting an indicated sound at the beginning, middle, or end of the word (e.g., say *game* without the /g/). The measure used for analyses was proportion correct. Cronbach's alpha internal consistency reliability for this sample was .72.

**Orthographic knowledge.** A two-alternative forced-choice task was designed using a subset of 20 items from the orthographic choice task of Olson, Forsberg, Wise, and Rack (1994). This new test includes 20 pairs of words, with each pair consisting of the correct spelling of a target word and a homophonic nonword spelling (e.g., *rain-rane* and *tertle-turtle*). Children indicated which of the two alternatives was the correct spelling for a target word used in a sentence. Because the word pairs are phonologically similar but orthographically dissimilar, children had to use word-specific knowledge of spelling and could not rely on phonological information to complete this task. The dependent measure was proportion correct. Cronbach's alpha internal consistency reliability for this sample was .62.

### Design

**Practice phase.** Children were randomly assigned to one of two groups, one of which received repeated practice reading a list of 40 words and one of which received repeated practice spelling the same list of 40 words.

The 40 practice words were chosen from eight different orthographic families. There were eight groups of five words that all



shared an orthographic rime (e.g., *chore*, *shore*, *snore*, *core*, and *sore*). All words within a family had consistent phonological-orthographic correspondences; that is, all orthographic patterns within a family of words sounded the same. Practice words are presented in Appendix A.

In the reading group, words were presented one at a time on 4-in.  $\times$  6-in. flashcards, with words that shared an orthographic rime presented one after another. Previous studies have indicated that learning is facilitated with a blocked presentation (e.g., Levy, 2001). The shared orthographic pattern within a family of words was printed in red to make the orthographic pattern maximally salient. Order of families within a trial and order of words within a family were randomly ordered across participants and, within participants, across trials. Participants were asked to read each word as accurately as possible. If a participant did not respond within 5 s, or responded incorrectly, the experimenter provided whole-word feedback by reading the word aloud. The dependent measure was the proportion of words read correctly. One practice trial was defined as reading through the list of 40 words once. Participants completed four trials a day over 4 days for a total of 16 repetitions.

The same list of 40 words was used for spelling practice. Words were presented orally by the experimenter. Words that shared an orthographic rime pattern were presented one after another. Participants printed the spelling of each word on a sheet of lined notepaper using a No. 2 pencil. On completion of each word, participants were visually presented with the word using the same flashcards as used in the reading practice condition. Participants self-corrected their spellings following each word. The dependent measure was the proportion of words spelled correctly before self-correction. All other aspects of spelling practice were the same as in the reading practice condition.

It is important to note that because children were able to view their written responses and the feedback provided for self-correction, it is possible that children in this condition were also reading the words, although they were not explicitly instructed to do so. This potential confound is inevitable when using written spelling; thus, this practice condition might more aptly be called a "spelling + reading" condition.

**Postpractice phase.** During the postpractice test phase, generalization within a skill and both word-specific and general transfer across skill were measured. Word-specific transfer across skill was assessed by having the participants either read or spell the 40 practice words, doing the opposite of what they had done during practice. To examine generalization within skill and generalization across skill, two different lists (List A and List B in Appendix B) of 64 new words were constructed. Each of these lists was made up of 32 new words that contained a practiced orthographic rime pattern (i.e., transfer words) and 32 new words that did not contain a practiced orthographic rime pattern (i.e., control words). All eight orthographic families were tested, with four new words from each family in the list. Orthographic patterns were not repeated across the two lists for the new words with unpracticed patterns to ensure that children did not become familiar with the patterns through testing. The lists were equal in the number of new words with a two-consonant cluster beginning and a one-consonant beginning. Word lengths were balanced across the two lists.

New words with or without practiced orthographic rime units and words within an orthographic family were randomly distrib-

uted throughout the list of 64 words. Participants read one list of 64 new words and spelled the other list of 64 new words. Lists were counterbalanced across participants such that half of the spellers and half of the readers spelled List A and read List B, and the other half of spellers and readers read List A and spelled List B. In addition, half of the spellers and half of the readers did the spelling first, and half of the spellers and half of the readers did the reading first. Presentation of test material was the same as presentation of material during the practice phase, except that no feedback was provided during the test phase. The proportion of words read or spelled correctly was the dependent measure. Table 1 presents an illustration of the design of the study for one child in the spelling group. Please note that only one orthographic family is illustrated in this example.

### Procedure

The selection and pretest measures (WJ-III reading, spelling, vocabulary, orthographic choice, and Test of Auditory Analysis Skills) were individually administered in one 30-min session. All other sessions involved individual testing as well. Over the next 4 days, participants either read or spelled the list of 40 practice words four times each day, in sessions lasting between 10 and 30 min. Spelling sessions generally lasted longer than did reading sessions. On the 5th day, posttest measures were administered in one 30-min session. All children began with the word-specific transfer task, followed by either the within-skill or across-skill generalization task. Order of presentation of these two tasks was counterbalanced across participants.

## Results

### Group Comparisons

Analyses of means on the pretest measures, presented in Table 2, were conducted to ensure that there were no preexisting differences between the two practice groups on any of the reading-related skills. A series of independent-sample *t* tests with practice group as the independent factor indicated that there were no reliable differences between groups on word identification,  $t(39) = -0.42$ ; spelling,  $t(39) = -1.76$ ; vocabulary,  $t(39) = 0.41$ ; orthographic knowledge,  $t(39) = -0.023$ ; or phonological processing

Table 1  
Illustration of Study Design

	Generalization within skill		Transfer across skill		
	Spell new with unit	Spell new without unit	Read practice words	Read new with unit	Read new without unit
Spelling practice					
<b>chick</b>	<b>slick</b>	droop	<b>chick</b>	<b>click</b>	truck
<b>stick</b>	<b>thick</b>	stoop	<b>stick</b>	<b>trick</b>	pluck
<b>brick</b>	<b>flick</b>	troop	<b>brick</b>	<b>prick</b>	stuck
<b>lick</b>	<b>pick</b>	hoop	<b>lick</b>	<b>sick</b>	puck
<b>tick</b>			<b>tick</b>		

*Note.* Boldface indicates when the trained unit was present in the generalization words and does not indicate how the words were presented in the study.

Table 2  
Means and Standard Deviations on Pretest Measures

Practice group	Word Identification	Spelling	Vocabulary	TAAS	Orthographic choice
Readers ( <i>n</i> = 20)					
<i>M</i>	100.25	100.90	107.25	.62	.74
<i>SD</i>	7.86	6.30	7.65	.24	.28
Spellers ( <i>n</i> = 21)					
<i>M</i>	101.33	104.71	106.29	.69	.75
<i>SD</i>	8.69	7.51	7.59	.21	.33

Note. Word Identification, Spelling, and Vocabulary are standard scores ( $M = 100$ ,  $SD = 15$ ) from the Woodcock-Johnson Tests of Achievement—Third Edition; Test of Auditory Analysis Skills (TAAS) and Orthographic Choice are proportion correct.

skill (Test of Auditory Analysis Skills),  $t(39) = -0.92$ , all  $ps > .05$ . Therefore, any benefits in reading or spelling outcome measures following the practice phase cannot be attributed to any confounds of preexisting differences in these reading-related skills.

### Practice Results

Level of performance across practice trials is presented in Figure 1. It is important to remember that the 40 words practiced were the same for both groups. What differed was the type of practice. For the readers, the mean proportion of words read correctly across trials is presented. For the spellers, the mean proportion of words spelled correctly across trials is presented. As illustrated in Figure 1, performance of both groups improved over trials. This observation was supported by a  $2 \times 16$  mixed analysis of variance with group (speller or reader) as the between-subject factor and

trial (1–16) as the within-subject factor. A significant main effect of trial was found,  $F(15, 585) = 41.70$ ,  $p < .001$ ,  $\eta^2 = .52$ . No other effects were significant, illustrating that both groups improved equally over trials. Although the readers were consistently slightly more accurate than the spellers, groups did not reliably differ in overall accuracy.

### Generalization Within Skill

Table 3 presents the proportion of new words with and without practiced rime patterns read correctly by the readers and spelled correctly by the spellers. A  $2 \times 2$  mixed analysis of variance, with practice group (reader or speller) as the between-subject factor and word type (practiced or unpracticed) as the within-subject factor, revealed a significant main effect of word type,  $F(1, 39) = 60.00$ ,  $p < .001$ ,  $\eta^2 = .61$ , and a significant interaction between word type and group,  $F(1, 39) = 5.53$ ,  $p < .05$ ,  $\eta^2 = .12$ . Simple effects analyses indicated that both groups showed generalization within skill, as readers read more new words with practiced rime patterns than new words with unpracticed rime patterns, and spellers spelled more new words with practiced rime patterns than new words with unpracticed rime patterns. However, generalization was not the same for both groups. Although both groups were equally accurate with the new words with practiced units, the difference between new words with and without practiced units was greater for the spellers (25%) than for the readers (13%),  $t(39) = -2.35$ ,  $p < .05$ ,  $d = 0.75$ , reflecting the spellers' less accurate performance with new words with unpracticed units.

### Transfer Across Skill

Transfer across skill was tested in two ways. First, I examined word-specific transfer to determine whether practice reading (or spelling) a specific word facilitates spelling (or reading) that

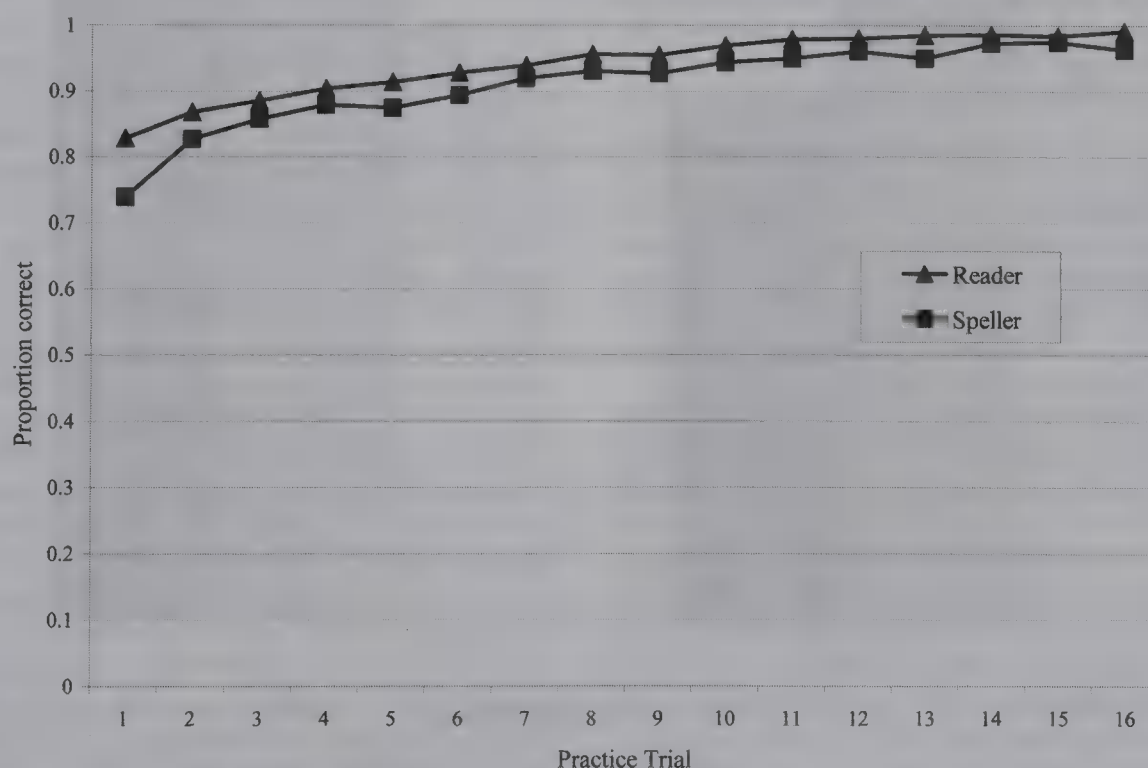


Figure 1. Proportion of words read and spelled correctly during the practice phase.



Table 3  
*Generalization Within Skill: Proportion of New Words With and Without Practiced Patterns Read Correctly by the Readers and Spelled Correctly by the Spellers*

Practice group	New words with practiced units	New words with unpracticed units
Readers ( $n = 20$ )		
<i>M</i>	.79	.65
<i>SD</i>	.20	.24
Spellers ( $n = 21$ )		
<i>M</i>	.74	.49
<i>SD</i>	.19	.23

specific word on a later encounter. Second, I examined generalization across skill—that is, whether practice in reading a set of words with a shared orthographic rime pattern benefits spelling unpracticed words that contain that same rime pattern, and vice versa.

*Word-specific transfer.* Figure 2 presents the proportion of the 40 practice words read and spelled correctly by each group. For the readers, the figure represents the proportion of practice words read correctly on the last trial of practice (Trial 16) and the proportion of practice words spelled correctly on the word-specific transfer task administered after practice. For the spellers, the figure represents the proportion of practice words spelled correctly on the last trial of practice (Trial 16) and the proportion of practice words

read correctly on the word-specific transfer task administered following practice. A  $2 \times 2$  mixed analysis of variance with group (readers or spellers) as the between-subjects factor and task (spelling or reading) as the within-subject factor revealed significant main effects of group,  $F(1, 39) = 27.88, p < .001, \eta^2 = .42$ , and task,  $F(1, 39) = 57.67, p < .001, \eta^2 = .60$ , and a significant interaction between group and task,  $F(1, 39) = 56.15, p < .001, \eta^2 = .59$ . Figure 2 clearly portrays the nature of this interaction. Following spelling practice, the spellers were equally able to read or spell the practice words, illustrating complete transfer from spelling to reading. However, the same is not true for reading practice, where there was less complete transfer to spelling. Practice reading a word did not necessarily ensure that a word could be spelled correctly.

*Generalization across skill.* Table 4 presents results relevant to generalization across skill by illustrating the proportion of practiced and unpracticed words spelled correctly by the readers and the proportion of practiced and unpracticed words read correctly by the spellers on each of the postpractice transfer tasks. I conducted a  $2 \times 3$  mixed ANOVA with practice group (readers or spellers) as the between-subjects factor and word type (practiced, new with practiced pattern, or new without practiced pattern) as the within-subject factor to examine transfer across skill. A significant main effect of word type,  $F(2, 78) = 36.06, p < .001, \eta^2 = .48$ , indicated that both groups were showing transfer across skill. Bonferroni post hoc analyses indicated that both groups were most accurate with the actual words they had practiced, followed

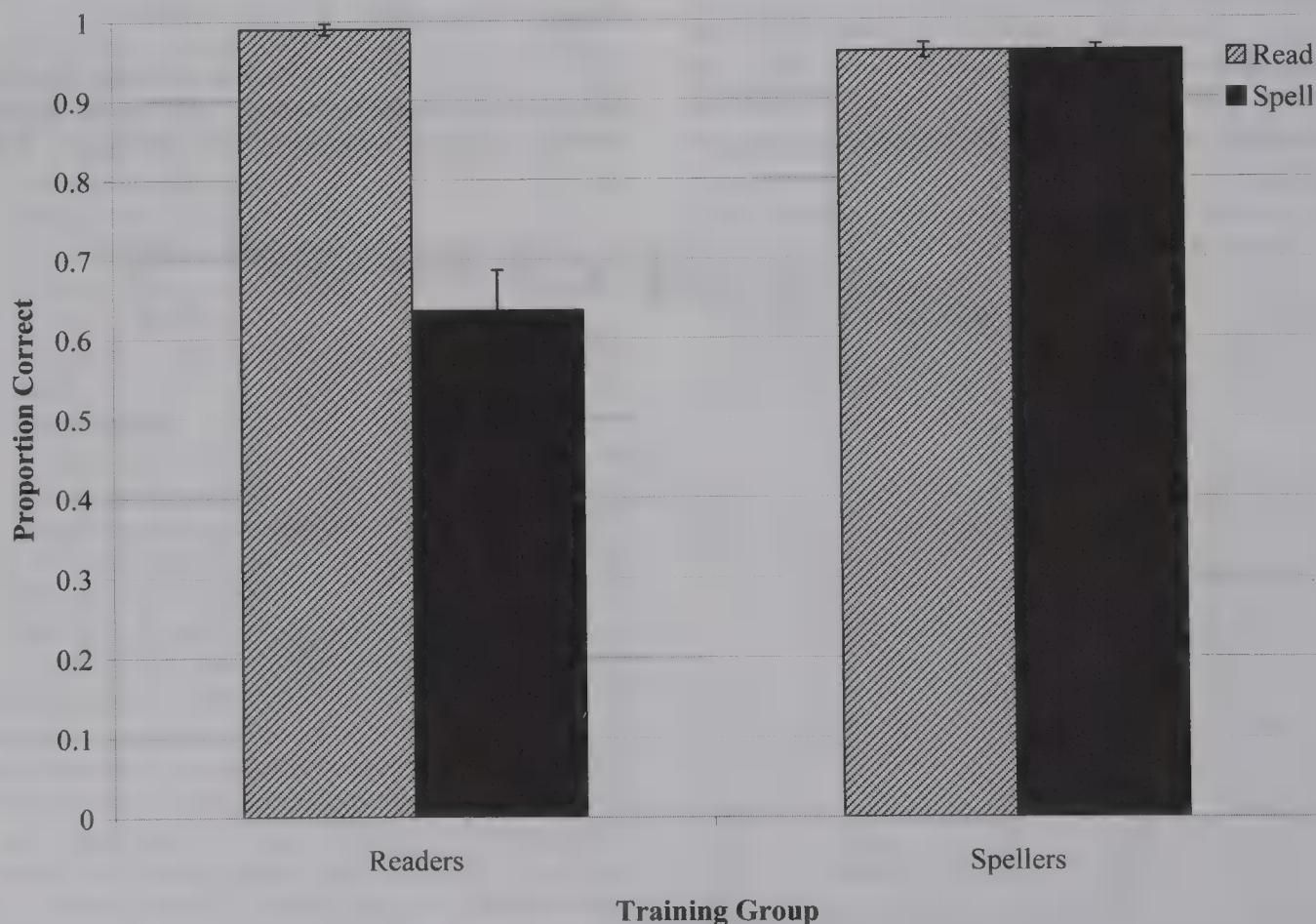


Figure 2. Word-specific transfer across skill: Proportion of practice words read and spelled correctly by both readers and spellers.

Table 4  
*Generalization Across Skill: Proportion of Practiced Words and New Words With and Without Practiced Patterns Spelled Correctly by the Readers and Read Correctly by the Spellers*

Practice group	Practiced words	New words with practiced units	New words with unpracticed units
Readers ( $n = 20$ )			
<i>M</i>	.64	.54	.40
<i>SD</i>	.23	.25	.22
Spellers ( $n = 21$ )			
<i>M</i>	.96	.87	.80
<i>SD</i>	.04	.10	.15

by new words with practiced patterns, which was more accurate than new words with unpracticed patterns. Thus, for both groups, word-specific transfer was greatest, but generalization across skill also occurred.

A significant main effect of group,  $F(1, 39) = 50.23, p < .001$ ,  $\eta^2 = .56$ , indicated that the spellers accurately read more words overall than the readers accurately spelled. The type of practice did not significantly interact with word type.

### Discussion

The main goal of the present study was to examine whether transfer can occur between reading and spelling in typically developing readers. Although research has suggested that the orthographic representations necessary to support fluent reading are established through practice and/or exposure to the printed word, how different types of practice contribute to the development of orthographic knowledge has not been fully investigated. The current results provide strong support for transfer between reading and spelling and suggest that spelling practice may be particularly valuable in setting up orthographic representations that can be used for reading.

Both word-specific and general transfer were found between reading and spelling. That word-specific transfer was greatest suggests that both reading and spelling practice set up word-specific representations that can be used across reading and spelling. However, generalization to new instances across skill also occurs, suggesting that children are able to abstract out the common orthographic pattern to aid in reading or spelling new words. Thus, in addition to building word-specific orthographic representations, children may also be increasing their general orthographic knowledge through reading and spelling. A few explanations are available to account for the generalization across skill found here. Several studies have indicated that children are able to read (e.g., Ehri & Robbins, 1992; Goswami, 1986) and spell (e.g., Bosse et al., 2003; Campbell, 1985) unfamiliar words through analogy to words already known. This account suggests that through reading and spelling practice, readers establish word-specific orthographic knowledge that can then be compared with unfamiliar words.

However, it is also possible that children establish more general knowledge of orthographic consistencies and patterns. For example, Fletcher-Flinn, Thompson, and colleagues (e.g., Fletcher-Flinn & Thompson, 2004; Thompson et al., 1996) have suggested that through practice with words children implicitly acquire knowledge of the relations between orthography and phonology

that are common to words the child has experienced. These "induced sublexical relations" can be at the grapheme-phoneme level or at the level of sequences of graphemes-phonemes. Whereas both explanations account for how a child may read an unfamiliar word, the explanations differ in terms of the nature of the orthographic representation or knowledge that is required. For analogy, the reader must have stored in memory a representation of a word neighbor to the unfamiliar word (e.g., *clock* must be stored to read *block*). In the sublexical relation account, the representation consists of relations formed between grapheme(s) and phoneme(s) that are acquired from all words in the child's reading vocabulary. In the present study, generalization may have occurred across skill as a result of children using analogies from the words they had practiced or through knowledge gained during practice through the induction of sublexical relations. The mechanism through which generalization from reading to spelling and from spelling to reading occurs needs to be further investigated.

Although transfer was found both from reading to spelling and from spelling to reading, the amount of transfer was unequal. At the end of the practice phase, spellers were equally able to read or spell the practice words, illustrating complete transfer from spelling to reading. However, the same was not found after reading practice. Practice reading a word did not necessarily ensure that word could be spelled correctly. This result suggests that the quality of the representations established through reading and spelling practice may differ. Spelling appears to set up a more detailed and accurate representation, as transfer from spelling to reading was significantly greater than transfer from reading to spelling.

The level of generalization within skill found in the present study also suggests that spelling practice may set up more detailed and accurate orthographic representation than practice reading. Although generalization within skill was found after both reading and spelling practice, the gains were significantly greater after spelling practice. That is, after spelling practice, children showed an increase in spelling accuracy between new words with trained and untrained patterns that was greater than the increase in reading accuracy after reading practice. It is important to point out, however, that there was more opportunity for improvement for the spellers, as accuracy for new words with untrained patterns was much lower than for the readers. Together, these results are consistent with theories that suggest that spelling requires a more specified orthographic representation than reading (Perfetti, 1997) and highlight the importance of spelling instruction in the early school years. Such instruction and practice can benefit developing reading skills and spelling skills.

The generalization within skill found in this study was greater than the amount of generalization reported in previous studies using similar methodologies of reading practice (Berends & Reitsma, 2006; Levy, 2001; Thaler et al., 2004). This difference most likely lies in the samples of children used in the various studies. Previous studies examined generalization within poor readers, whereas the present study used average readers. There is evidence to suggest that generalization after practice such as this will differ across skill level. For example, in the reading-by-analogy literature, Ehri and Robbins (1992) demonstrated that readers first had to have a minimal level of decoding skill before they were able to use analogy to generalize from familiar words to unfamiliar words. The goal of the present study was to examine whether transfer and



generalization could occur across reading and spelling in typically developing children. Because the present study indicates that transfer can occur, even with words that are familiar to children, future studies should examine whether practice spelling is a viable method to improve the reading skills of children with reading disabilities.

Last, it is important to note that the present study was a training study to examine whether transfer *can* occur. Despite this fact, there are similarities between the conditions of the current study and what may occur in a naturalistic setting. At the beginning of the practice sessions, children were quite accurate in reading and spelling the practice words, indicating that they already had familiarity with the words. As noted previously, practice with words results in increasingly accurate orthographic representations. In everyday reading and spelling, children are practicing words that are familiar to them, continually improving the quality of the orthographic representations. In addition, spelling a word typically takes longer than reading a word and involves a motor response. The extra time spent on spelling or the additional motor cues may help account for the greater benefits of spelling practice. Also, the spelling practice condition may more appropriately have been labeled as a "spelling + reading" condition, as it involved both reading and spelling. When children were provided with visual feedback to correct their spelling, they may also have been reading words. However, this also occurs when children are spelling in their everyday lives. They spell the word. Afterward, it is visible for them to read. Although familiarity, time spent practicing, the motor response, and the confound of spelling involving reading are important considerations for future research to help identify which specific components of spelling practice are most influential in setting up detailed orthographic representations, from an educational perspective the results still provide strong support for the benefit of spelling practice in addition to reading practice.

In conclusion, the present study illustrates that transfer between reading and spelling occurs in both directions. Although repeated reading and spelling of isolated words is perhaps not the most practical route through which children establish orthographic representations, the findings from the current study do inform how we should teach our children to read and spell. Reading and spelling programs should be coordinated to benefit students maximally. In addition, transfer from spelling to reading was greater than transfer from reading to spelling, suggesting that spelling practice may establish a more detailed and accurate orthographic representation than does practice reading. This finding indicates that instruction in spelling may be particularly valuable in the early school years to aid in developing reading skills. Finally, results suggest that spelling practice should be further investigated to determine whether such practice is a viable method to improve the reading skills of children with reading disabilities.

## References

- Adams, M. J. (1990). *Beginning to read: Thinking and learning about print*. Cambridge, MA: MIT Press.
- Barker, T. A., Torgesen, J. K., & Wagner, R. K. (1992). The role of orthographic processing skills on five different reading tasks. *Reading Research Quarterly*, 27, 334–345.
- Beck, I., & McKeown, M. (1991). Conditions of vocabulary acquisition. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.),

- Handbook of reading research: Volume 2* (pp. 789–814). Hillsdale, NJ: Erlbaum.
- Berends, I. E., & Reitsma, P. (2006). Remediation of fluency: Word specific or generalized training. *Reading and Writing: An Interdisciplinary Journal*, 19, 221–234.
- Berninger, V. W., Vaughan, K., Abbott, R. D., Brooks, A., Abbott, S. P., Rogan, L., et al. (1998). Early intervention for spelling problems: Teaching functional spelling units of varying size with a multiple-connections framework. *Journal of Educational Psychology*, 90, 587–605.
- Bosman, A. M. T., & van Orden, G. C. (1997). Why spelling is more difficult than reading. In C. A. Perfetti, L. Rieben, & M. Fayol (Eds.), *Learning to spell: Research, practice and theory across languages* (pp. 173–194). Mahwah, NJ: Erlbaum.
- Bosse, M., Valdois, S., & Tainturier, M. (2003). Analogy without priming in early spelling development. *Reading and Writing: An Interdisciplinary Journal*, 16, 693–716.
- Bradley, L., & Bryant, P. E. (1983, February 3). Categorizing sounds and learning to read: A causal connection. *Nature*, 301, 419–421.
- Campbell, R. (1985). When children write nonwords to dictation. *Journal of Experimental Child Psychology*, 40, 133–151.
- Cunningham, A. E., Perry, K. E., Stanovich, K. E., & Share, D. L. (2002). Orthographic learning during reading: Examining the role of self-teaching. *Journal of Experimental Child Psychology*, 82, 185–199.
- Ehri, L. C. (1980). The development of orthographic images. In U. Frith (Ed.), *Cognitive processes in spelling* (pp. 311–338). London: Academic Press.
- Ehri, L. C. (1997). Learning to read and learning to spell are one and the same, almost. In C. A. Perfetti, L. Rieben, & M. Fayol (Eds.), *Learning to spell: Research, practice and theory across languages* (pp. 237–269). Mahwah, NJ: Erlbaum.
- Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading*, 9, 167–188.
- Ehri, L. C., & Robbins, C. (1992). Beginners need some decoding skill to read words by analogy. *Reading Research Quarterly*, 27, 12–26.
- Ehri, L. C., & Wilce, L. (1986). The influence of spellings on speech: Are alveolar flaps /d/ or /t/? In D. Yaden & S. Templeton (Eds.), *Metalinguistic awareness and beginning literacy* (pp. 101–114). Portsmouth, NH: Heinemann.
- Ehri, L. C., & Wilce, L. (1987). Does learning to spell help beginners learn to read words? *Reading Research Quarterly*, 22, 47–65.
- Fletcher-Flinn, C. M., & Thompson, G. B. (2004). Mechanisms of implicit lexicalized phonological recoding used concurrently with underdeveloped explicit letter sound skills in both precocious and normal reading development. *Cognition*, 90, 303–355.
- Goswami, U. (1986). Children's use of analogy in learning to read: A developmental study. *Journal of Experimental Child Psychology*, 42, 73–83.
- Hulme, C. (1981). *Reading retardation and multi-sensory teaching*. London: Routledge & Kegan Paul.
- Laberge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6, 293–323.
- Levy, B. A. (2001). Moving the bottom: Improving reading fluency. In M. Wolf (Ed.), *Dyslexia, fluency, and the brain* (pp. 357–379). Timonium, MD: York Press.
- Meyer, M. S., & Felton, R. H. (1999). Repeated reading to enhance fluency: Old approaches and new directions. *Annals of Dyslexia*, 49, 283–306.
- Nathan, R. G., & Stanovich, K. E. (1991). The causes and consequences of differences in reading fluency. *Theory Into Practice*, 30, 176–184.
- Olson, R., Forsberg, H., Wise, B., & Rack, J. (1994). Measurement of word recognition, orthographic, and phonological skills. In G. R. Lyons (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues* (pp. 243–277). Baltimore: Paul H. Brookes.

- Perfetti, C. A. (1992). The representation problem in reading acquisition. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 145–174). Hillsdale, NJ: Erlbaum.
- Perfetti, C. A. (1997). The psycholinguistics of spelling and reading. In C. A. Perfetti, L. Rieban, & M. Fayol (Eds.), *Learning to spell: Research, practice and theory across languages* (pp. 21–38). Mahwah, NJ: Erlbaum.
- Perfetti, C. A., & Hart, L. (2001). The lexical basis of comprehension skill. In D. S. Gorfein (Ed.), *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity* (pp. 67–96). Washington, DC: American Psychological Association.
- Roberts, K., & Ehri, L. C. (1983). Effects of two types of letter rehearsal on skilled and less skilled beginning readers' word memory. *Contemporary Educational Psychology*, 8, 375–390.
- Rosner, J. (1979). *Test of auditory analysis skills*. Novato, CA: Academic Therapy.
- Share, D. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, 55, 151–218.
- Share, D. (2004). Orthographic learning at a glance: On the time course and developmental onset of self-teaching. *Journal of Experimental Child Psychology*, 87, 267–298.
- Snowling, M. J. (2000). Language and literacy skills: Who is at risk and why? In D. V. M. Bishop & L. B. Leonard (Eds.), *Speech and language impairments in children: Causes, characteristics, intervention and outcome* (pp. 245–259). New York: Psychology Press.
- Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, 21, 360–406.
- Statistics Canada. (2002). *2001 community profiles*. Retrieved December 9, 2007, from <http://www12.statcan.ca/english/profil01/CP01/Index.cfm?Lang=E>
- Thaler, V., Ebner, E. M., Wimmer, H., & Landerl, K. (2004). Training reading fluency in dysfluent readers with high reading accuracy: Word specific effects but low transfer to untrained words. *Annals of Dyslexia*, 54, 89–113.
- Thompson, G. B., Cottrell, D. S., & Fletcher-Flinn, C. M. (1996). Sub-lexical orthographic-phonological relations early in the acquisition of reading: The knowledge sources account. *Journal of Experimental Child Psychology*, 62, 190–222.
- Woodcock, R. W., McGrew, K. S., & Mathers, N. (2001). *Woodcock-Johnson III Tests of Achievement*. Itasca, IL: Riverside.

## Appendix A

### Practice Words

think	clock
stink	block
pink	dock
link	mock
blink	flock
skunk	shore
trunk	tore
sunk	chore
punk	score
bunk	pore
plump	chick
lump	sick
bump	brick
grump	trick
chump	pick
chest	black
guest	rack
pest	stack
nest	hack
zest	sack

(Appendixes continue)



## Appendix B

## Generalization and Transfer Lists

List A		List B	
drink	bless	brink	bloom
slink	dress	clink	groom
rink	less	sink	doom
wink	mess	kink	boom
stunk	bride	plunk	luck
chunk	ride	spunk	pluck
dunk	wide	junk	cluck
hunk	tide	gunk	puck
clump	bring	slump	bleed
pump	sing	hump	freed
dump	thing	jump	seed
stump	wing	thump	need
crest	place	lest	bake
best	race	jest	cake
test	lace	west	flake
rest	space	vest	brake
shock	broke	stock	start
lock	spoke	sock	part
jock	smoke	frock	mart
smock	woke	rock	dart
swore	blade	store	dire
more	grade	sore	fire
lore	made	core	wire
wore	wade	yore	hire
flick	book	click	stoop
tick	brook	pick	droop
prick	hook	lick	troop
kick	cook	thick	hoop
crack	drive	track	chill
back	hive	tack	frill
flack	live	smack	still
lack	dive	pack	pill

Received September 13, 2007

Revision received March 26, 2008

Accepted April 2, 2008 ■

# A Meta-Analysis of Single Subject Design Writing Intervention Research

Leslie Ann Rogers and Steve Graham  
Vanderbilt University

There is considerable concern that students do not develop the writing skills needed for school, occupational, or personal success. A frequent explanation for this is that schools do not do a good job of teaching this complex skill. A recent meta-analysis of true- and quasi-experimental writing intervention research (S. Graham & D. Perin, 2007a) addressed this issue by identifying effective instructional writing practices. The current review extends this earlier work by conducting a meta-analysis of single subject design writing intervention studies. The authors located 88 single subject design studies where it was possible to calculate an effect size. They calculated an average effect size for treatments that were tested in 4 or more studies, using a similar outcome measure in each study. This resulted in the identification of 9 writing treatments that were supported as effective. These were strategy instruction for planning/drafting, teaching grammar and usage, goal setting for productivity, strategy instruction for editing, writing with a word processor, reinforcing specific writing outcomes, use of prewriting activities, teaching sentence construction skills, and strategy instruction for paragraph writing.

**Keywords:** writing, composition, meta-analysis, instruction, single subject design

According to the most recent National Assessment of Educational Progress (Persky, Daane, & Jin, 2003), many children do not learn to write well enough to meet classroom writing demands. The writing of 2 out of every 3 students in Grades 4, 8, and 12 was below grade-level proficiency. Concerns about writing are not limited to elementary and secondary schools, however, as college instructors estimated that 50% of high school graduates are not prepared for college-level writing demands (Achieve, Inc., 2005). Moreover, American businesses spend \$3.1 billion annually for writing remediation (National Commission on Writing, 2004). In 2003, these problems led the National Commission on Writing to conclude that the writing of students in the United States “is not what it should be” (p. 7).

Students who do not learn to write well are at a disadvantage. At school, weaker writers are less likely than their more skilled peers to use writing to support and extend learning in content classrooms (Graham & Perin, 2007a). Their grades are more likely to suffer, particularly in classes where writing is the main tool (e.g., via tests and papers) for assessing progress (Graham, 2006b). Their chances of attending college are reduced, as writing is now used to evaluate many applicants’ qualifications. At work, writing has become a gateway for employment and promotion (see reports by the National Commission on Writing, 2004, 2005). Employees in business and government, for instance, are expected to produce written documentation, visual presentations, memoranda, technical re-

ports, and electronic messages. Socially, adults who do not write well may not be able to participate fully in civic life, as e-mail and text messaging have progressively supplanted telephones as a primary means for communicating. On a personal level, people use writing to explore who they are, to combat loneliness, and to chronicle their experiences. Writing about one’s feelings and experiences, for example, is beneficial psychologically and physiologically (see Smyth, 1998, for a meta-analysis of this research).

Why do so many students not write well enough to meet grade level demands? One possible reason is that schools do not do an adequate job of teaching this complex skill. The National Commission on Writing (2003) charged that this is the most neglected of the three *Rs* in the American classroom and offered the following recommendations: double time students spend writing, assess their writing progress, use technology to advance the learning and teaching of writing, and better prepare teachers to teach writing. The impact of these recommendations is likely to be reduced if teachers do not use effective instructional practices.

## Need for a Meta-Analysis of Writing Interventions Tested Via Single Subject Design Studies

A useful approach for identifying effective writing practices is to conduct systematic reviews of writing intervention research. Since the mid-1980s there have been a number of meta-analyses of the writing intervention literature (e.g., Hillocks, 1986; Graham & Perin, 2007a, 2007b). With meta-analysis, an effect size is computed for each empirical study investigating a specific treatment and then is averaged across studies to provide a summary statistic on the intervention’s effectiveness (Lipsey & Wilson, 2001). Most of the meta-analyses of writing intervention research have focused on treatments tested in true- and quasi-experimental studies (Bangert-Drowns, 1993; Bangert-Drowns, Hurley, & Wilkinson, 2004; Goldring, Russell, & Cook, 2003; Graham, 2006a; Graham & Harris, 2003; Graham & Perin, 2007a, 2007b; Hillocks, 1986).

---

*Editor's Note.* Tom Scruggs served as the action editor for this article.—KRH

---

Leslie Ann Rogers and Steve Graham, Department of Special Education, Vanderbilt University.

Correspondence concerning this article should be addressed to Steve Graham, Peabody 328, 230 Appleton Place, Vanderbilt University, Nashville, TN 37203. E-mail: steve.graham@vanderbilt.edu



However, two of the meta-analyses also computed effect sizes for treatments tested via single subject designs (Graham, 2006a; Graham & Harris, 2003). These two reviews only examined a single treatment: teaching strategies for planning/revising. Consequently, the primary purpose of this article was to conduct a more extensive meta-analysis of single subject design writing interventions in order to identify effective writing practices for students in Grades 1–12.

Like true-experimental studies, single subject designs can be used to test whether a treatment is responsible for observed changes in performance. As Horner et al. (2005) noted, “Single-subject design is experimental rather than correlational or descriptive, and its purpose is to document causal or functional relationships between independent and dependent variables” (p. 166). Major threats to internal validity are controlled through within- and between-subjects comparisons, and external validity is enhanced through systematic replication. In single subject design studies, each participant serves as her/his own control, with performance prior to as well as during and/or after intervention repeatedly measured to establish performance patterns before treatment and comparison of performance patterns across experimental phases (e.g., baseline versus treatment). To establish experimental control, the independent variable or treatment is actively manipulated to determine its effects on the dependent measure(s) (Horner et al., 2005).

One manipulation (reversal) for establishing experimental control involves introduction and withdrawal of the treatment (e.g., a stable baseline pattern of performance is established, followed by introduction of treatment to determine whether it influences the dependent measure[s], followed by withdrawal of treatment to determine whether performance returns to or near baseline levels, followed by reintroduction of treatment to determine whether performance is again influenced).

A second manipulation (multiple baselines) involves the staggered introduction of the treatment. This can involve both within- and between-subjects comparisons. For example, after researchers establish a stable baseline pattern of performance for each participating student, treatment is implemented with one student to determine whether it influences the students’ performance in a predictable fashion. If the instructed student demonstrates the predicted pattern of change, while uninstructed students’ performance remains unchanged from baseline, then the treatment is implemented with the next student to determine whether the pattern described above is replicated. This systematic delay in introduction of the treatment continues until all students receive instruction.

With both of these manipulations, experimental control is established only if performance on the dependent measures is stable during each experimental phase (e.g., baseline and treatment) and there is no trend in the pattern of baseline performance in the direction predicted by the intervention. In addition, it is generally agreed that experimental control is not established until there are at least three demonstrations that the manipulation had the predicted impact. Thus, experimental control is established when the predicted covariation between the introduction of the treatment and changes in the dependent measure(s) are demonstrated through at least three demonstrations in an experiment.

In contrast to true- and quasi-experimental designs where the focus is on group performance, single subject design examines the

effectiveness of a treatment at the individual level (single subject design studies, however, typically include more than a single student). Considerable emphasis is placed on fully describing the participants, the context in which the investigation took place, and factors that influence participants’ performance prior to intervention (Horner et al., 2005). Such rich descriptions set the stage for determining the boundaries of generalization, but a basic tenet in single subject design research is that the generality of a treatment is not established in a single study. Instead, external validity is established by systematically replicating effects across multiple participants, locations, and researchers.

### Why a Meta-Analysis of Single Subject Design Studies Is Important

There are three reasons why it is advantageous to conduct a systematic, empirical, and full review of single subject design writing intervention research. First, the meta-analyses of true- and quasi-experimental investigations of writing interventions have identified only 12 interventions that improve the writing of elementary and secondary students (see Bangert-Drowns, 1993; Bangert-Drowns et al., 2004; Goldring et al., 2003; Graham, 2006a; Graham & Harris, 2003; Graham & Olinghouse, *in press*; Graham & Perin, 2007a, 2007b; Hillocks, 1986). These include (ranked according to the magnitude of their impact): teaching strategies for planning, revising, and editing; teaching written summarization; having students help each other carry out specific writing processes such as planning, drafting, or revising/editing; establishing specific goals for students’ writing; teaching handwriting, spelling, or typing to students; using word processing as a medium for writing; teaching students how to write more complex sentences through sentence combining activities; engaging students in inquiry as a means for developing ideas for writing; encouraging students to engage in prewriting activities to gather and organize possible writing content; establishing a process approach to writing; using writing as a tool for content learning; and having students study and emulate models of good writing. While the identification of these 12 treatments is an important accomplishment, these practices do not cover all aspects of learning to write. A meta-analysis of single subject design writing interventions has the potential to broaden current evidence-based recommendations.

Second, a meta-analysis of single subject design studies also has the potential to strengthen, undermine, or nuance the trust that can be placed in one or more of the 12 writing treatments identified as effective above. For example, converging evidence from true- and quasi-experimental studies as well as single subject design research would bolster the claim that a treatment was or was not effective, whereas conflicting evidence would undermine such a claim or foster a more nuanced conclusion. The need for additional evidence is also supported by the fact that only 4 of the 12 treatments identified above (teaching writing strategies, process writing approach, word processing, and using writing to learn) were based on 10 or more studies.

Third, most of the true- and quasi-experimental writing intervention research has been conducted with students representing the full range of writing ability in a typical classroom (Graham & Perin, 2007a). The only exceptions to this involve strategy instruction, word processing, and setting goals for students’ writing. In



contrast, single subject design studies often involve students' experiencing difficulty. Consequently, a meta-analysis of single subject writing intervention research has the potential to increase the number of identified evidence-based practices for struggling writers.

It is also important to establish that there is currently no comprehensive meta-analysis of single subject writing intervention studies. Graham and colleagues (Graham, 2006a; Graham & Harris, 2003) conducted meta-analyses of single subject design studies examining strategy instruction in planning, revising, and/or editing. The first of these reviews (Graham & Harris, 2003) only focused on a specific model of strategy instruction, whereas the second review (Graham, 2006a) concentrated on all single subject design studies in this area. In both reviews, teaching writing strategies had a positive impact on writing, providing additional support to the findings from true- and quasi-experimental studies (e.g., Graham & Perin, 2007a).

### The Current Meta-Analysis

The meta-analysis reported in this article draws on but greatly extends the two previous meta-analyses of single subject strategy instruction research in writing (Graham, 2006a; Graham & Harris, 2003). We not only broadened the review to include all writing practices tested via single subject design but conducted a broader search than Graham (2006a) or Graham and Harris (2003), resulting in the identification of some studies not included in their review. To reduce the likelihood of only identifying studies where positive effects were obtained (i.e., nonsignificant findings are rarely published in peer-refereed journals), we searched as broadly as possible, including studies published in journals as well as dissertations, theses, and book chapters.

The primary research question guiding this review was, Which writing practices tested via single subject design procedures are effective with students in Grades 1–12? Consistent with Graham and Perin (2007a), no conclusions were drawn about a treatment unless there were at least four studies with a conceptually similar outcome measure assessing its impact. Furthermore, we assessed the quality of each single subject design study included in this review, using quality indicators developed by Horner et al. (2005). This allowed us to identify strengths and weaknesses in the current body of single subject writing intervention research and temper conclusions about the effectiveness of a treatment based on the quality of the research.

The theoretical bases for many of the writing treatments in this review were grounded in behavioral theory. This is not surprising, as single subject design methodology grew out of behavioral work in the 1960s (Horner et al., 2005). Nevertheless, some treatments, such as strategy instruction, were influenced by cognitive (Hayes, 2000) and/or behavioral models, whereas the theoretical underpinnings of other treatments were unstated (e.g., word processing). Consequently, we draw no claims about the validity of specific theories.

### Method

#### *Location and Selection of Studies*

The strategies that we used to locate and select studies for this meta-analysis were influenced by five factors. First, studies were

included that involved students in Grades 1–12. This differed from Hillocks's (1986) comprehensive meta-analyses of true- and quasi-experimental writing intervention research that focused on students in Grade 3 to college, and Graham and Perin (2007a, 2007b), who concentrated on Grades 4–12. As noted earlier, our goal was to identify effective writing treatments for both elementary and secondary students.

Second, we included studies that were conducted with students attending regular public schools, private schools, alternative schools, summer programs, clinics, and residential centers. We cast a broader net than Graham and Perin (2007a, 2007b), who did not include studies conducted in special schools for persons with disabilities (e.g., residential centers), as we were especially interested in identifying writing practices that would be effective with struggling writers, including those with disabilities (we did not purposefully exclude any type of disability).

Third, we included only single subject designs with mechanisms for establishing experimental control (i.e., demonstrating a functional relationship between the independent and dependent variable). This included reversal and multiple baseline designs (described in the introduction) as well as alternating treatment and changing criterion designs. Alternating treatment designs involve the rapid alternation of two or more distinct treatments (with each treatment presented the same number of times) in a counterbalanced fashion to determine their impact on a single outcome measure. The treatments are alternated rapidly to reduce the possibility of carryover effects, and treatments are counterbalanced to eliminate order effects. We only included alternating treatment design studies that included a baseline (this is a recommended practice but is not absolutely essential). Baseline is established before introduction of the alternating treatments or by including it as one of the alternating treatment conditions (i.e., a no-treatment condition). Experimental control is established when a treatment repeatedly produces the predicted change in behavior in reference to baseline (whether this is a baseline established before alternating treatments or a no-treatment control included as one of the alternating treatments).

With a changing criterion design, the desired outcome during treatment is gradually and systematically increased or decreased for a specific behavior. After establishing a stable baseline pattern of performance, the researcher introduces the treatment (e.g., reinforcement) but divided into subphases, with two or more subphases requiring increased changes (if the desire is to improve the behavior; e.g., "Write 20 words," "Write 30 words") and at least one subphase requiring decreased changes (e.g., "Now write only 15 words"). Such changes are meant to gradually move the student toward the desired terminal goal (e.g., "Write 50 words") while demonstrating that it is treatment, and not maturation, that causes the change. Experimental control is established if the student's behavior matches or exceeds the predetermined criterion specified in each subphase of treatment.

Fourth, studies were included if they provided the data needed to calculate the effect size, percentage of nonoverlapping data (PND; this measure is described more fully later). This required that the data at each assessment point was provided either in graph or tabular form. If a study did not have such data for at least one writing measure, it was not included in this review. In addition, we did not include studies where baseline data were not collected. Nor did we include studies that focused solely on the teaching of



handwriting or spelling, as they typically do not examine whether such instruction impacts students' actual writing (Graham, 1999).

Fifth, we searched broadly to identify possible studies. This involved locating peer-refereed and nonrefereed studies from a variety of sources, including studies in prior meta-analyses (i.e., Graham, 2006a; Graham & Harris, 2003), journals, theses and dissertations, conference proceedings, and books. We searched broadly, going beyond published peer-refereed articles, to reduce the possibility of bias, as studies are rarely published in peer-refereed journals when they do not obtain positive treatment results.

A number of databases were searched during January 2007 for relevant studies, including ERIC, PsycINFO, ProQuest, Education Abstracts (i.e., Education Full Text), and Dissertation Abstracts. We ran multiple searches in these databases, pairing writing, composition, and narrative with the following terms: *dictation, genre, genre and instruction, goal setting, grammar, inquiry, mechanics, models, peer collaboration, peer planning, peer revising, peers, planning, revising, pre-writing activities, process writing, reinforcement, self-evaluation, self-monitoring, sentence combining, sentence construction, speech synthesis, spell checkers, strategy instruction, summary instruction, summary strategies, summary writing, technology, usage and mechanics, word processing, word processor, writer's workshop, and writing assessment*. These terms were also included in the search by Graham and Perin (2007a).

These same databases were searched to identify single subject design writing intervention research conducted by 11 prominent researchers in this area (e.g., K. Ballard, T. Glynn, and C. MacArthur). We also conducted a hand search of the following journals that frequently publish single subject design research: *Education and Treatment of Children, Educational Psychologist, Exceptional Children, Journal of Applied Behavior Analysis, Journal of Behavioral Education, Journal of Learning Disabilities, Journal of Special Education Technology, Learning Disabilities & Practice, Learning Disability Quarterly, and Remedial & Special Education*. Finally, the reference lists for all obtained articles were searched to identify relevant studies.

Using these search mechanisms, we collected 119 documents and found 88 studies that were suitable for this review. The most common reason for excluding a study was that it did not include a writing outcome measure. This occurred in 11 instances. Nine studies were eliminated because they did not apply a design where experimental control could be established, whereas 5 studies were excluded because they contained no baseline phase. Finally, 3 studies were omitted because the same data were published elsewhere, 2 did not provide the data needed to calculate PND, and 1 study included a combined writing/reading intervention.

### *Categorizing Studies Into Treatment Conditions*

First, each study was read and placed into a treatment identified in advance. These treatments corresponded to the descriptors used in the electronic searches. These categories were supplemented by two additional treatments common in behavioral research: reinforcement and direct instruction. Studies that did not fit neatly into one of these preidentified categories were held apart until all studies were read and sorted. At this point, the studies in each preidentified treatment were reread to determine whether the in-

tervention in each study represented the same general treatment. If this was not the case, they were placed with the studies that were not classified during the initial reading. All of the studies in the unclassified pile were read again, resulting in the construction of new treatments. The studies in these new treatments and any preidentified treatment where an additional study was placed were again reread to determine whether each intervention represented the same general treatment. As this process took place, we refined some initial treatment categories and eliminated others (when no studies tested that treatment).

At least one or more studies examined the effectiveness of the following 20 treatments (ordered from most to least frequently studied): strategy instruction for planning/drafting (25 studies), self-monitoring (8 studies), goal setting for productivity (7 studies), reinforcement (6 studies), prewriting activities (5 studies), sentence construction (5 studies), strategy instruction editing (5 studies), strategy instruction paragraph construction (5 studies), word processing (5 studies), teaching grammar/usage (4 studies), feedback on writing (4 studies), strategy instruction other (4 studies), strategy instruction revising (2 studies), direct instruction of a broad array of skills (2 studies), goal setting for grammar/sentence construction (2 studies), word processing plus (2 studies; these studies included word processing plus an additional support, for example, text read aloud by the computer), dialogue journals (1 study), direct teaching of self-regulation strategies (1 study), repeated writing (1 study), and verbal encouragement (1 study).

The most common treatment involved teaching strategies for carrying out specific aspects of the writing process. Instead of lumping these into a single category, we separated them into different categories depending on the processes emphasized. Most of these studies focused on teaching a genre-specific writing strategy for planning and drafting papers (stories, persuasive essays, and expository text), and this formed a single category. All of the planning/drafting strategies in these 25 investigations were taught using the Self-Regulated Strategy Development model (SRSD; see Harris & Graham, 1996, for a description of this teaching approach).

The other strategy instructional categories that we created concentrated on teaching strategies for writing paragraphs, editing only, or revising/editing. Since these investigations emphasized different processes (from each other and the strategy for planning/drafting studies), we created a category for each. Finally, a small number of strategy instructional studies did not fit neatly into any of these categories and were placed in a strategy instruction other category. Studies in this category included investigations where strategy instruction was combined with other practices (such as a token reinforcement system; Boyer, 1990), students were taught a summary writing strategy (Nelson, Smith, & Dodd, 1992), or the study did not assess students' independent use of the strategy (Li, 2000). Our decision to create multiple strategy treatments differed from Graham and Perin (2007a, 2007b), where all strategy studies were included in the same category.

We also differed from Graham and Perin (2007a, 2007b) in that we created two treatment categories for goal setting. The purpose of the goals in these two categories differed, as one set of goals was aimed at increasing productivity, and the other focused on grammatical correctness. It should be noted that we created two word processing treatment categories. This included word processing, which examined the effects of word processing and software



commonly bundled into such a program (e.g., spell checkers) as well as word processing plus, which included word processing plus programs like word prediction (Handley-More, Deitz, Billingsley, & Coggins, 2003) and speech synthesis programs that read written text aloud (Channon, 2004).

As noted earlier, we calculated only a summary statistic (mean PND) for treatments that included four or more studies with a conceptually similar outcome measure. Table 1 presents a definition for the 10 treatment categories that met this condition. We decided to adopt this criterion, as it was used by Hillocks (1986) in his seminal review and Graham and Perin (2007a, 2007b) in their more recent review. If reviewers use similar criterion, it is easier to make valid comparisons across reviews. We recognize, however, that small sample sizes are less reliable and must be interpreted more cautiously than a summary statistic based on a larger number of studies.

The decision to stress a conceptually similar outcome measure in computing a summary statistic also has precedence in the two prior comprehensive reviews of true- and quasi-experimental research (Graham & Perin, 2007a, 2007b; Hillocks, 1986). In those reviews, the focus was only on the outcome measure of writing quality (it was assumed that there would be less noise or error in the analyses, if the outcome measures were conceptually similar). With the single subject design studies reported here, researchers typically graphed just the measure that was most directly associated with the intervention (e.g., if the intervention was meant to increase writing output, then this variable was graphed). Thus, there was no single outcome measure for most of the studies in this review. For half of the treatments described in Table 1, there were at least four or more studies that graphed writing productivity (number of words and sentences). These treatments were goal setting for productivity, reinforcement, self-monitoring, strategy instruction for planning/drafting, and word processing. Two treatments (prewriting and strategy instruction for planning/drafting) met the four study criteria for quality of writing (as measured with holistic rating scales; see Graham & Perin, 2007a), whereas an-

other two treatments (strategy instruction for planning/drafting and strategy instruction for paragraph construction) met it for elements (basic structural elements of a genre or writing task, such as story parts). Sentence construction, strategy instruction for editing, and teaching grammar/usage had four or more studies that graphed number of complete sentences, errors corrected, and grammar errors, respectively.

### *Coding of Study Features*

*Descriptive information.* For each study that met the established criteria, the following 10 pieces of information were collected and coded: type of design (i.e., withdrawal design, multiple-baseline design, alternating treatment design, and changing criteria design), number of participants, type of writers (i.e., full range of writers in a typical classroom, above average writers only, average writers only, struggling writers only, and English language learners), grade of study participants, age of study participants, disability status of each participant (i.e., attention-deficit/hyperactivity disorder [ADHD], behavior disorder, emotional and behavior disorder, emotionally disturbed, learning disability, mild language delay, mild mental retardation, other health impairment, orthopedic impairment, speech and language delay, visual impairment, and Section 504), race/ethnicity of participants (i.e., African American, American Indian, Caucasian, Hispanic, Arab, or Other), geographical location of study (i.e., suburban, rural, or urban), person providing instruction (i.e., teacher or member of research team), and written description of independent variable (coded as one of the 20 treatment categories established in the previous section).

*Quality indicators.* In addition to this descriptive information, each study was evaluated to see whether it met 11 specific indicators of study quality proposed by Horner et al. (2005) for single subject design research. Some of these quality indicators were based in part on the descriptive information collected and coded above, but others required the collection of additional information.

Table 1  
*Summary of Evidence-Based Practices for Teaching Writing*

Treatment category	Definition
Strategy instruction: Planning/drafting	Students were taught strategies for planning, drafting, and/or revising stories, persuasive essays, and/or expository essays. This included modeling of the strategy and guided practice to facilitate independent use of it.
Teaching grammar/usage	Students were taught grammar and usage skills via well-sequenced and highly focused instruction.
Goal setting for productivity	Students were provided or set an explicit goal to write more.
Strategy instruction: Editing	Students were taught a strategy for editing their papers. This included modeling of the strategy and guided practice to facilitate independent use of it.
Word processing	Students used word processing as their medium for writing.
Reinforcement	Students received praise, public recognition, or tangible items contingent on writing improvement(s).
Prewriting activities	Students used graphic organizers, including story maps and outlines, for generating ideas prior to writing.
Sentence construction	Students were taught sentence skills through either sentence combining or a strategy for writing sentences.
Strategy instruction: Paragraph construction	Students were taught a strategy for writing a paragraph. This included modeling of the strategy and guided practice to facilitate independent use of it.
Self-monitoring	Students self-monitored either their on-task behavior, writing productivity, or writing quality, and the results of their performance was displayed.



The score for each indicator ranged from 0 to 1 (with a score of 1 indicating that the quality indicator was met), and all scores for a study were summed to obtain a total quality score.

The quality indicator, participant description, was based on five or six pieces of information, depending on whether students with disabilities were included in the study. A study received a score of 1 on this indicator, if information was provided on participants' (a) ages or grades, (b) gender, (c) socioeconomic status, (d) ethnicity, (e) writing achievement at the start of the study, and when appropriate (f) disability status. The score for this indicator was calculated by assigning 1 point for each of these items if they were reported, and then dividing by either 5 or 6, depending on whether students with disabilities were included in the study.

A score of 1 was awarded if the researcher described the procedures for determining how students were selected to participate in the study (this needed to be done in enough detail so that another researcher could replicate these procedures). A score of 1 was also assigned if the location where instruction took place was described in enough detail so that it could easily be visualized. For each indicator, a score of 0 was awarded if the criterion was not met.

For the quality indicator, operationally defining dependent variables, we examined whether the procedures for scoring dependent variables (that were graphed) were described in enough detail to allow other researchers to use these measures. If all dependent measures were defined in this fashion, a score of 1 was assigned. A score of 0 was assigned if no dependent measures met this criterion. A score of .50 was awarded if at least one of the dependent variables met this criterion.

The quality indicator, interrater reliability of graphed dependent variables, was assessed by calculating the proportion of dependent measures that were scored reliably. If reliability for all dependent measures equaled or exceeded .60, a score of 1 was awarded. A score of 0 was awarded if none of the dependent variables met this criterion. If the reliability for some but not all dependent variables was .60 or higher, then the number of dependent measures meeting criterion was divided by total number of dependent variables. Researchers used different means to calculate reliability (ranging from percentage agreement to correlation coefficients).

For the quality indicator, multiple baseline data points, we examined each baseline to determine whether it had three or more data points. A score of 1 was assigned if all baselines met this criterion, whereas a score of 0 was assigned if no baselines met it. If some but not all baselines met this criterion, then the proportion that met it was calculated. The exact same procedures were used to score the quality indicator, multiple intervention data points.

The remaining four quality indicators (treatment description, fidelity of treatment, testing procedure descriptions, and social validity) received a score of either 1 or 0. A score of 1 was awarded for each of the following: a treatment was described in enough detail so that it could be applied by others, the researcher collected and reported data that demonstrated fidelity of treatment (treatment was delivered as intended), enough detail (e.g., directions) was provided so that other researchers would be able to administer the writing assessments, and data on the social validity of the treatment were collected (as suggested by Kennedy, 2002, this could include questioning students, teachers, or others about the social validity of the treatment or a demonstration that treat-

ment gains were maintained for 3 or more weeks following instruction).

In addition to these 11 quality indicators, we also examined whether experimental control was established. This is critical to establishing a functional relation between the independent and dependent variables. Experimental control was established if there were at least three demonstrations in a study of an experimental effect that met these criteria: (a) the last three data points in baseline established a stable pattern of behavior (i.e., the last three data points fluctuated by no more than 20% from the highest score of the three data points), (b) there was no trend in baseline performance in the direction predicted by the treatment, (c) and treatment had the predicted impact on behavior. To determine whether treatment had the predicted impact on behavior (letter *c* above), 50% or more of the treatment data points had to exceed the strongest baseline score.

In determining whether experimental control was established, it was necessary to customize the third criterion above to the four different single subject designs included in this review. For withdrawal designs, treatment had the predicted impact on behavior when 50% or more of the data points in the first introduction of the treatment ( $B_1$ ) surpassed the strongest initial baseline score ( $A_1$ ) and strongest score when treatment was withdrawn and the second baseline was initiated ( $B_2$ ). For multiple baseline designs, 50% or more of the treatment data points had to exceed the strongest baseline score, and there could be no corresponding improvement or trend toward improvement for each baseline that had not received treatment yet. For alternating treatment designs, 50% or more of the treatment scores for a specific treatment had to exceed the strongest score from baseline. This included the strongest score from either a baseline established before treatments were alternated and/or baseline scores collected as an alternating condition (i.e., no treatment). For changing criterion designs, all of the treatment data points from the first subphase of treatment had to fall within 20% of the stated goal for that subphase, with this same criterion applied to subsequent subphases.

**Reliability.** To establish the reliability of scoring procedures, 20% of the studies ( $n = 18$ ) were randomly selected and rescored. The second reader was trained and scored all 18 studies independent of Leslie Rogers. Interrater reliability was established for each scoring procedure separately. The scoring procedures were reliable, as percentage of agreement ranged from 89% to 100%, with a mean percentage of agreement across all variables of 96%.

### *Calculation of Effect Sizes*

In order to assess overall treatment effects, we employed a nonparametric approach to meta-analysis and calculated the percentage of nonoverlapping data points between baseline and treatment phases: referred to as PND (Scruggs, Mastropieri, & Casto, 1987). PND is the percentage of data points in treatment that represent an improvement over the most positive value obtained during baseline. As recommended by Scruggs, Mastropieri, Cook, and Escobar (1986), PND was not calculated when ceiling or floor levels were evident during baseline (depending on the intended direction of the outcome variable), as there was no room for treatment effects to be realized.

For each study, we calculated a separate PND for each measure that provided the needed information. PND for a specific measure



in a study was calculated by first obtaining PND for each baseline treatment comparison (or changing condition) and then calculating an average PND across all relevant changing conditions for that measure. When possible, we calculated a PND for treatment, maintenance, and generalization for each measure, as we were interested not just in immediate effects but in the impact of the treatment over time as well as transfer effects to other situations. A PND calculated for treatment involved comparing scores during treatment, immediately following treatment, or both to the strongest baseline score for each changing condition in the study. A PND calculated for maintenance involved comparing a student's scores collected 3 weeks or more after treatment ended to the strongest baseline score for each changing condition in the study. A PND for generalization (typically to another genre or setting) involved comparing a student's generalization scores during or after treatment to the strongest generalization score during baseline for each changing condition in the study.

When there were four or more studies of a treatment that included a conceptually similar outcome measure, we calculated a mean, median, and range of PNDs for that measure across studies (confidence intervals were not calculated as PND lacks a known sampling distribution; Parker, Hagan-Burke, & Vannest, 2007). When possible, we calculated these three indices (mean, median, and range) for treatment, maintenance, and generalization. PND was interpreted using criteria proposed by Scruggs et al. (1986): PND greater than 90% is a large effect, PND between 70.1% and 90% is a moderate effect, PND between 50.1% and 70% is a low or small effect, and PND 50% or below is classified as not effective.

To avoid overinflating the importance of a single study, it is recommended that a single effect size be calculated for each study (Lipsey & Wilson, 2001). Although we computed more than one PND for many studies, this basic concept was followed here. For all but one of the treatments, only one effect size from each study was used to calculate summary PNDs. In all of these instances, we were only able to calculate a summary PND for a single measure for each treatment. The only exception involved the treatment of strategy instruction for planning/drafting. Almost all of the studies in this category had a conceptually similar outcome measure that was graphed (i.e., elements), but there were also enough studies that conjointly graphed productivity and writing quality that we were able to calculate a separate summary PND for these measures too. We decided to present a summary PND for these measures separately, as it made little sense to develop a single summary PND combining all three measures or simply ignore two of them. The summary PND for a measure was never based on more than one effect size from a single study, however. The same was true when a summary PND was computed for treatment, maintenance, or generalization.

PND was calculated for all studies by Leslie Rogers, and reliability was established by a second rater who calculated PND for 18 randomly selected studies. Reliability was .99.

Why did we decide to use PND over other methods for calculating effect sizes for single subject design studies? First, as Scruggs and Mastropieri (1998) noted, single subject design effect sizes created by subtracting mean baseline performance from mean treatment performance and dividing by baseline (or the pooled) standard deviation (much like Cohen's *d*) do not take into account the within-subject nature of the data and can result in effect sizes

that are idiosyncratic and meaningless (e.g., when we have used this method for SRSD single subject design studies in writing, the effect sizes are typically 3.0 and higher; they average just over 1.0 in true- and quasi-experimental studies; Graham & Perin, 2007a). Other alternative methods, such as the regression effect size method (which examines the proportion of student score variance explained by phase differences), also possess limitations that made them less suitable for our analysis (see Scruggs & Mastropieri, 1998, for a discussion of different approaches). For example, the parametric assumptions (normality, equal variance, and serial independence of data) underlying techniques such as the regression effect method are not commonly met in single subject design studies (PND is not bound by these parametric assumptions; see Parker et al., 2007). Second, we would have had to eliminate many studies from this review if the regression approach were applied, as it was not possible to determine the exact scores for each data point from graphs in many studies (it was possible, however, to compute PND, as it is only necessary to establish that one score was greater than another). Third, PND was the most commonly used measure in previous meta-analyses of single subject writing intervention research; Graham, 2006a; Graham & Harris, 2003), making it easier to compare findings across reviews.

## Results

Table 2 presents basic information about the individual studies that tested the effectiveness of the 10 treatments that included four or more studies with a conceptually similar outcome measure (see Table 1 for a description of these treatments). This includes information about study design, participants (number, type of writer, grade, age, and race/ethnicity), geographic location, overall quality of study (based on the 11 quality indicators), whether experimental control was established, as well as the PNDs for specific measures. The Appendix includes information about studies that were included in treatments with three or less studies with a conceptually similar outcome measure. Table 3 presents summary PNDs (mean, median, and range) for each of the treatment categories included in Table 2, whereas Table 4 presents summary PNDs for the treatment, strategy instruction for planning/drafting, by type of genre (story versus expository).

Table 5 presents the average quality score for studies in each of the 10 treatments included in Table 2. The total quality score was the sum of the scores for the 11 quality indicators (e.g., participant description and participant selection). For each treatment, we also report the percentage of studies where each quality indicator was met. This same information is presented cumulatively for all studies included in the 10 treatments, providing a general indication of the quality of the single subject design intervention studies on which our findings are based.

Before examining individual treatments, we first offer some comments on the overall quality of single subject design writing intervention studies. As can be seen in Table 5, almost all studies quantified dependent variables, established reliability of variables, collected three or more data points in baseline as well as intervention, and adequately described the intervention. However, participant description and selection were only adequately described in about one half of the studies, and an adequate description of the

(text continues on page 892)



Table 2  
Writing Instruction Treatments That Included Four or More Studies

Study	Design	N	Writer type	Grade (Age)	Disability	Race/ethnicity	Location	Instr.	Quality score (0-11)	PND TX	PND POST	PND MAINT	PND GEN
Strategy instruction (planning/drafting)													
Stories Adkins (2005)	MBD	3	SW	2 to 3 (7&9)	EBD	AA	SUB	RES	10.00		100% <sup>a</sup> (Elements) 100% <sup>a</sup> (Production) 100% <sup>a</sup> (Quality)	100% (Elements) 100% (Production) 67% (Quality)	100% (Elements) Personal Narratives 100% (Production) Personal Narratives 67% (Quality) Personal Narratives
Albertson: Study 1 (1998)	MBD	4	SW	5 to 8 (10-13)	ADHD (n = 1)	NR	NR	RES	6.67	95% (Elements) 68% <sup>a</sup> (Production) 91% <sup>a</sup> (Grammar)			
Albertson: Study 2 (1998)	MBD	2	SW	6 (11-12)	—	NR	NR	Teacher	6.60	100% (Elements) 92% (Production) 100% (Grammar)			
Albertson & Billingsley (1997)	MBD	2	AVG	6 (12)	Gifted	NR	NR	RES	7.08	67% (Elements) 78% (Production)			
Danoff et al. (1993)	MBD	6	FR	4 to 5	LD (n = 3)	Asian (n = 1) H (n = 1) C (n = 4)	SUB	Teacher	8.67		100% <sup>a</sup> (Elements)	100% (Elements)	100% (Elements) Setting
Lane et al. (2007)	MBD	6	SW	2 (7-8)	—	AA (n = 2) C (n = 4)	RURAL	RES	10.00		100% <sup>a</sup> (Elements)	100% (Elements)	
Lienemann et al. (2006)	MBD	6	SW	2 (7-8)	ADHD (n = 1) LD (n = 1) OI (n = 1)	AA (n = 1) C (n = 4) H (n = 1)	RURAL	RES	10.00		97% <sup>a</sup> (Elements)	83% (Elements)	67% (Elements) Reading
Reid & Lienemann (2006)	MBD	3	SW	3 to 4 (9-10)	ADHD (n = 1)	C	RURAL	RES	11.00		100% (Elements) 100% (Production)	100% (Elements) 100% (Production)	
Saddler et al. (2004)	MBD	6	SW	2 (7-8)	MLD (n = 1) SLI (n = 1)	AA	SUB	RES	10.00		100% <sup>a</sup> (Elements)	88% (Elements)	82% (Elements) Personal Narratives
Germain (2004)	MBD	10	AVG	4 (9-10)	—	AA (n = 2) H (n = 1) C (n = 7)	SUB	RES	7.00	95% <sup>a</sup> (Quality)	100% <sup>a</sup> (Quality)		

Table 2 (continued)

Study	Design	N	Writer type	Grade (Age)	Disability	Race/ethnicity	Location	Instr.	Quality score (0-11)	PND TX	PND POST	PND MAINT	PND GEN
Saddler (2006)	MBD	6	SW	2 (7)	LD	AA	URBAN	RES	9.00		100% <sup>a</sup> (Elements) 100% <sup>a</sup> (Production) 100% <sup>a</sup> (Quality)	100% (Elements) 83% (Production) 92% (Quality)	
Saddler & Asaro (2007)	MBD	6	SW	2 (7)	LD	AA (n = 4) C (n = 2)	URBAN	RES	8.00		100% <sup>a</sup> (Elements) 100% <sup>a</sup> (Quality)		
Harris & Graham (1985)	MBD	2	SW	— (12)	LD	NR	SUB	RES	8.67		96% <sup>a</sup> (Production)	33% (Production)	100% (Production) Setting 89% (Grammar) Less Dominant Language
Korducki (2001)	MBD	8	ESOL	5 (10-11)	LD (n = 4)	H	URBAN	RES	9.83		92% (Grammar)	75% (Grammar)	
Guzel-Ozmen (2006)	MBD	4	SW	— (13-17)	MMR	Arab	NR	RES	7.83		100% <sup>a</sup> (Structure Element)	100% (Structure Element)	
Troia et al. (1999)	MBD	3	SW	5 (10-11)	LD	AA (n = 1) C (n = 2)	SUB	RES	9.66		100% <sup>a</sup> (Elements)	67% (Elements)	67% (Elements) Persuasive
Persuasive essays De La Paz & Graham (1996)	MBD	3	SW	5	LD	AA (n = 2) C (n = 1)	SUB	RES	8.83		100% <sup>a</sup> (Elements)	100% (Elements)	
Graham & Harris (1989)	MBD	3	SW	6 (12)	LD	NR	SUB	RES	8.67		100% <sup>a</sup> (Elements)	89% (Elements)	100% (Elements) Setting 89% (Elements) Story
Graham et al. (1992)	MBD	4	SW	5 (11-12)	LD	NR	URBAN	RES	8.67		100% <sup>a</sup> (Elements)	100% (Elements)	
Lienemann (2006)	MBD	4	SW	4 to 5 (9-11)	LD (n = 1)	C	RURAL	RES	10.00		100% <sup>a</sup> (Elements) 100% <sup>a</sup> (Production) 89% <sup>a</sup> (Elements)	100% (Elements) 100% (Production) 100% (Elements)	
Mason & Shriner (2007)	MBD	6	SW	2 to 5 (8-12)	EBD (n = 4) LD (n = 1) 504 (n = 1)	NR	NR	RES	8.75				
Sexton et al. (1998)	MBD	6	SW	5 to 6 (10-12)	LD	AA (n = 5) C (n = 1)	SUB	RES	8.00		68% <sup>a</sup> (Elements)	25% (Elements)	100% (Elements) Setting
Expository essays De La Paz (2001)	MBD	3	SW	7 to 8	ADD (n = 1) SLI (n = 2)	C	SUB	Teacher	9.00		100% <sup>a</sup> (Elements)	100% (Elements)	

(table continues)



Table 2 (continued)

Study	Design	Writer type	Grade (Age)	Disability	Race/ ethnicity	Location	Instr.	Quality score (0-11)	PND TX	PND POST	PND MAINT	PND GEN
De La Paz (1999)	MBD	22 FR	7 to 8	LD (n = 6)	AA (n = 2) C (n = 20)	SUB	Teacher	10.00		95% <sup>a</sup> (Elements) 78% <sup>a</sup> (Production) 97% <sup>a</sup> (Quality)	100% (Elements) 88% (Production) 100% (Quality)	
Mason et al. (2006)	MBD	9 SW	4 (9-10)	EBD (n = 1) LD SLI (n = 1)	AA (n = 1) C (n = 8)	URBAN	RES	9.83		94% <sup>a</sup> (Production)	100% (Production)	84% (Production) Reading
Campbell et al. (1991)	ABAB & MBD	3 SW	2 (9)	MMR (n = 1) LD (n = 2)	AA (n = 1) C (n = 2)	NR	Teaching grammar/usage Peers	9.00	79% (Grammar)		100% (Grammar)	77% (Grammar) Other Untrained Peers
Dowis (1991)	MBD	14 SW	5 to 6	LD	NR	SUB	Teacher	8.50	75% <sup>a</sup> (Grammar)	87% (Grammar)	89% (Grammar)	
Dowis & Schloss (1992)	MBD	4 SW	6 (12)	LD	AA (n = 1) C (n = 3)	SUB	Teacher	8.00	88% <sup>a</sup> (Grammar)		76% (Grammar)	
Herman et al. (1976)	MBD	9 SW	6	—	NR	NR	Teacher	7.20	88% <sup>a</sup> (Grammar)	83% (Grammar)		
Hopman & Glynn (1989)	CCD	4 SW	HS (13)	—	NR	Goal setting for productivity URBAN	RES	6.40	26% (Production)	63% (Production)		
Kastelen et al. (1984)	MBD	16 FR	8 (13)	—	NR	NR	Teacher	6.20	100% (Production)		100% (Production)	
Seabaugh & Schumaker (1994)	MBD	3 FR	9 to 12 (14-18)	LD (n = 2)	NR	NR	RES	10.83	60% (Production)			
Van Houten et al. (1974)	ABAB	55 FR	2 & 5	—	NR	NR	Teacher	6.40	94% <sup>a</sup> (Production)			
Van Houten & McKillop (1977)	ABAB	38 FR	10 & 11	—	NR	NR	Teacher	6.40	91% <sup>a</sup> (Production)			
Weygant (1981)	MBD	9 SW	4 to 5 (10-12)	LD	C	NR	RES	9.00	99% (Production)			
Wolfe (1997)	CCD	4 SW	2 to 3 (9)	LD	NR	URBAN	Teacher	9.50	63% <sup>a</sup> (Production)			
Beals (1983)	MBD	9 FR	10 (15-16)	LD (n = 3)	AA (n = 4) C (n = 5)	Strategy instruction (editing) URBAN	Teacher	7.75		71% (Errors Corrected)		
McNaughton et al. (1997)	MBD	3 SW	10 & 12 (16-18)	LD	NR	RURAL	RES	7.67	95% <sup>a</sup> (Errors Corrected)	100% (Errors Corrected)	100% (Errors Corrected)	

Table 2 (continued)

Study	Design	Writer N type	Grade (Age)	Disability	Race/ ethnicity	Location	Instr.	Quality score (0-11)	PND TX	PND POST	PND MAINT	PND GEN
Schmidt (1983)	MBD	7 SW	10 to 12 (14-16)	LD	C	SUB	RES	7.00		50% (Errors Corrected) 29% (Elements)	34% (Errors Corrected) 34% (Elements)	
Schumaker et al. (1982)	MBD	9 SW	8 to 12 (12-18)	LD	NR	NR	Teacher	6.34		100% (Errors Corrected) 100% (Errors Detected) 100% (Errors / Word) 100% <sup>a</sup> (Errors Corrected)		
Scott (1993)	MBD	4 SW	4 (10-12)	LD	AA (n = 3) C (n = 1)	URBAN	RES	10.92				
Brigman (1994)	ATD	6 SW	1 (6-7)	—	NR	Word processing SUB & RURAL	Teacher	9.90	63% <sup>a</sup> (Production)			
Christensen (1993)	MBD	12 SW	4	Gifted	C	URBAN	Teacher	5.00	100% (Production)			
Handley-More et al. (2003)	ATD	3 SW	4 to 5 (10-11)	LD	NR	NR	RES	5.50	29% (Production)			
Jones (1985)	ABAB	6 SW	4 (9)	—	AA (n = 3) C (n = 3)	RURAL	RES	7.60	86% <sup>a</sup> (Production) 62% <sup>a</sup> (Holistic Content) 14% (Holistic Form)			
Burnett (1984)	ABAB	10 SW	1-3 & 5	—	C	RURAL	Teacher	8.80	64% <sup>a</sup> (Quality)			
Ballard & Glynn (1975)	MBD	14 FR	3 (8-9)	—	NR	Reinforcement RURAL	Teacher	8.20	100% <sup>a</sup> (Production)			
Blasé-Maloney & Hopkins (1973)	MBD	14 NR	4 to 6	—	NR	NR	Teacher	6.20	100% <sup>a</sup> (Production)			
Blasé-Maloney et al. (1975)	MBD	19 NR	3	—	NR	URBAN	Teacher	6.20	100% (Production)			
Brigham et al. (1972)	MBD	13 SW	5	—	NR	NR	Teacher	6.40	84% (Production)			
Bording et al. (1984)	MBD	9 SW	(12-16)	BD (n = 5) MMR (n = 2) LD (n = 2)	C	NR	Teacher	8.83	57% (Grammar)			
Newstrom et al. (1999)	MBD	1 SW	9	EBD	NR	NR	NR	5.50	100% (Grammar)			
Campbell & Willis (1978)	MBD	32 FR	5 (10-12)	—	NR	URBAN	Teacher	7.60	86% <sup>a</sup> (Creativity)	67% <sup>a</sup> (Creativity)		

(table continues)



Table 2 (continued)

Study	Design	N	Writer type	Grade (Age)	Disability	Race/ethnicity	Location	Instr.	Quality score (0-11)	PND TX	PND POST	PND MAINT	PND GEN
Channon (2004)	ATD	7	SW	8 (14-15)	LD	H	URBAN	Prewriting activities Teacher	8.83	84% <sup>a</sup> (Quality)			
Channon (2004) Inspiration Draft: Builder	ATD	7	SW	8 (14-15)	LD	H	URBAN	Teacher	8.83	13% (Quality)			
Thanhouser (1994)	MBD	6	SW	12 (17-18)	LD	AA ( <i>n</i> = 1) C ( <i>n</i> = 5)	SUB	RES	8.83	44% (Quality) 64% (Production)			38% (Quality) Setting 58% (Production)
Zipprich (1995)	MBD	3	SW	3 to 5 (9-10)	LD	C ( <i>n</i> = 2) H ( <i>n</i> = 1)	NR	Teacher	7.83	59% (Quality) 9% (Production)	70% (Quality) 17% (Production)	50% (Quality) 0% (Production)	
Martin & Manno (1995)	MBD	3	SW	7 (13)	LD	AA ( <i>n</i> = 1) C ( <i>n</i> = 1) AA & C ( <i>n</i> = 1)	URBAN	RES	6.50	75% (Elements)			
Beals (1983)	MBD	9	FR	10 (15-16)	LD ( <i>n</i> = 3)	AA ( <i>n</i> = 4) C ( <i>n</i> = 5)	URBAN	Sentence construction Teacher	7.75	100% (Complete Sentences)	88% (Complete Sentences) 80% (Complicated Sentences)		
First (1994)	MBD	3	SW	— (11-13)	ED	C ( <i>n</i> = 1) H ( <i>n</i> = 1) H&AA ( <i>n</i> = 1)	NR	Teacher	4.00	80% (Complete Sentences)			
Johnson (2005)	MBD	36	FR	7 to 8 (11-13)	Autistic MMR LD VI OHI	C ( <i>n</i> = 11) H ( <i>n</i> = 25)	NR	Teacher	9.00	83% (Complete Sentences)	89% (Complete Sentences)		
Schmidt et al. (1988)	MBD	7	SW	10 to 12 (14-16)	LD	C	SUB	Teacher	7.00	78% (Complete Sentences)	50% (Complete Sentences)		
Eads (1991)	MBD	9	SW	6 to 7 (12-14)	LD	NR	RURAL	Teacher	5.67	89% (Complete Sentences)			
Moran et al. (1981, study 1)	MBD	3	SW	8 to 9 (14-16)	LD	NR	SUB	Strategy instruction (paragraph construction) Teacher	6.67	100% (Elements)	100% (Elements) Setting		
Moran et al. (1981, study 2)	MBD	5	SW	8 to 9 (13-16)	LD	NR	SUB	Teacher	7.70	100% (Elements)			
Sonntag & McLaughlin (1984)	MBD	6	FR	8 to 9	—	NR	RURAL	Teacher	5.40	100% (Elements)			
Wallace & Bott (1989)	MBD	4	SW	8	LD	NR	RURAL	RES	6.67	89% (Elements)			100% (Elements) Setting

Table 2 (continued)

Study	Design	Writer type	Grade (Age)	Disability	Race/ ethnicity	Location	Instr.	Quality score (0-11)	PND TX	PND POST	PND MAINT	PND GEN
Dowell et al. (1994)	MBD	3 SW	9 to 11 (14-16)	LD	NR	URBAN	Teacher	5.58	34% (Production) 4% (Quality)			
Goddard (1998)	ABAB	7 SW	5 to 6 (10-12)	LD	C	SUB	Self-monitoring RES	9.83	29% (Production)			15% (Vocab.)
Harris et al. (1994)	MBD	4 SW	5 to 6 (10-12)	LD	AA (n = 1) C (n = 3)	SUB	Teacher	7.83	43% (Production)			
Jackson (1994)	MBD	6 SW	- (11-13)	—	NR	NR	Teacher	6.40	82% <sup>a</sup> (Production)			
Moran (2004)	MBD	4 SW	2 (7-8)	ADHD (n = 2)	AA (n = 1) AI (n = 1) C (n = 2)	SUB	RES	11.00	23% (Production)			
Rumsey & Ballard (1985)	ABAB	7 SW	— (9-11)	—	NR	NR	Teacher	5.90	52% <sup>a</sup> (Production)			
Shimabukuro (1999)	MBD	3 SW	6 to 7 (12-13)	ADHD & LD	NR	NR	Teacher	3.67	91% <sup>a</sup> (Production) 75% (Accuracy)			
Wolfe (1997)	ABAB	4 SW	2 to 3 (9)	LD	NR	URBAN	RES	9.50	37% (Production)			
Montague & Leavell (1994)	MBD	9 SW	7 to 8 (12-14)	LD	AA (n = 1) C (n = 5) H (n = 3)	URBAN	RES	5.33	29% (Quality)			

*Note.* Dashes indicate that data were not reported. ATD = alternating treatment design; ABAB = withdrawal design; CCD = changing criteria design; MBD = multiple-baseline design; N = Number of participants on single subject graph(s); AVG = average; ESOL = English as a second language; NR = not reported; SW = struggling writer; FR = full range; ADHD = attention-deficit/hyperactivity disorder; BD = behavioral disorder; EBD = emotional and behavioral disorder; ED = emotionally disturbed; LD = learning disability; MLD = mild language delay; MMR = mild mental retardation; OHI = other health impairment; OI = orthopedic impairment; SLI = speech or language impairment; VI = visual impairment; 504 = student plan with goals and objectives related to writing; AA = African American; AI = American Indian; C = Caucasian; H = Hispanic; SUB = suburban. Instr. = instructor; RES = researcher; PND = percentage of nonoverlapping data; TX = during treatment; POST = immediately after treatment (less than 3 weeks after treatment ended); MAINT = maintenance (3 or more weeks after treatment ended); GEN = generalization (generalization to other settings or genres); Elements = basic parts of a genre or type of writing; Production = included number of words, ideas, and/or T-units, as well as number of assignments completed; Vocab = included number of different types of words.

<sup>a</sup> Experimental control established.



Table 3  
Summary PND Statistics for Writing Treatments With Four or More Studies

Treatment category	Measure	N	Grade range	Writer type	M	Mdn	Range
Strategy instruction: Planning/drafting	Elements	21	2-8	FR, SW	96% TX/POST	100% TX/POST	67%-100% TX/POST
	Elements	18	2-8	FR, SW	90% MAINT	100% MAINT	25%-100% MAINT
	Elements	4	2-3, 5-6	SW	85% GEN Genre	86% GEN Genre	67%-100% GEN Genre
	Production	10	2-8	FR, SW	91% TX/POST	95% TX/POST	68%-100% TX/POST
	Production	7	2-5, 7-8	FR, SW	86% MAINT	100% MAINT	33%-100% MAINT
Teaching grammar/usage	Quality	5	2-4, 7-8	FR, SW	99% POST	100% POST	97%-100% POST
	Grammar	4	2, 5-6	SW	83% TX	84% TX	75%-88% TX
	Production	7	2-5, 8-12	FR, SW	79% TX/POST	91% TX	26%-100% TX
Goal setting for productivity	Errors	5	4, 8-12	FR, SW	84% POST	100% POST	50%-100% POST
	Corrected						
Word processing	Production	4	1, 4-5	SW	70% TX	75% TX	29%-100% TX
Reinforcement	Production	4	3-6	FR, SW	96% TX	100% TX	84%-100% TX
Prewriting activities	Quality	4	3-5, 8, 12	SW	52% TX/POST	55% TX/POST	13%-84% TX/POST
Sentence construction	Complete sentences	5	6-8, 10-12	FR, SW	86% TX/POST	83% TX/POST	78%-100% TX/POST
	Elements	4	8-9	FR, SW	97% TX/POST	100% TX/POST	89%-100% TX/POST
Strategy instruction: Paragraph construction	Elements	4	8-9	FR, SW	97% TX/POST	100% TX/POST	89%-100% TX/POST
Self-monitoring	Production	7	2, 3, 5-7	SW	51% TX	43% TX	23%-91% TX

*Note.* Production includes number of words, ideas, and/or T-units, as well as number of assignments completed. N = number of studies; FR = full range; SW = struggling writer; PND = percentage of nonoverlapping data points; TX = scores during treatment; TX/POST = scores during treatment as well as immediately after treatment (less than 3 weeks after treatment ended); MAINT = maintenance (3 or more weeks after treatment ended); GEN = scores for generalization probes.

physical location of the study was only offered about one third of the time (these findings raise concerns about external validity). Likewise, fidelity of treatment was established in less than one half of the studies (possibly because this has not been stressed until recently) but adequate description of testing procedures and the establishment of social validity of the treatment occurred in about two third of the studies.

The findings for each of the 10 treatments in Table 1 are presented next. Each treatment is presented separately using a common format. First, information on studies testing the treatment is summarized. This is followed by the summary PND statistics (mean, median, and range) and interpreting these statistics in light of the quality of the studies on which they were based (the 11 quality indicators in Table 5 and the establishment of experimental control; see Table 2).

#### Strategy Instruction: Planning/Drafting

Twenty-five studies examined the effectiveness of teaching strategies for planning/drafting specific types of text (see Table 2). As noted earlier, all of these investigations used the SRSD model (Harris & Graham, 1996) to teach these strategies. A common component in most of the planning/drafting strategies taught was that students used specific features of the target genre (e.g., characters, location, time, goals of characters, actions, ending, and reactions for stories) to help them generate and organize possible writing ideas. Planning/drafting strategies focused on stories (e.g., Adkins, 2005; Albertson, 1998; Albertson & Billingsley, 1997), persuasive text (e.g., De La Paz & Graham, 1997; Lienemann, 2006; Mason & Shriner, 2008), and expository essays (e.g., De La

Table 4  
Summary PND Statistics for Strategy Instruction Using Planning/Drafting for Stories and Expository Writing

Measure	N	Grade range	Writer type	M	Mdn	Range
Stories						
Elements	13	2-8	FR, SW	97% TX/POST	100% TX/POST	67%-100% TX/POST
Elements	9	2-5	FR, SW	93% MAINT	100% MAINT	67%-100% MAINT
Production	7	2-8	FR, SW	91% TX/POST	96% TX/POST	68%-100% TX/POST
Production	4	2-5	SW	79% MAINT	92% MAINT	33%-100% MAINT
Expository						
Elements	8	2-8	FR, SW	94% POST	100% POST	68%-100% POST
Elements	8	2-8	FR, SW	89% MAINT	100% MAINT	25%-100% MAINT

*Note.* Elements = basic parts of a genre or type of writing; Production includes number of words, ideas, and/or T-units, as well as number of assignments completed. N: number of studies; FR = full range; SW = struggling writer; PND = percentage of nonoverlapping data points; TX = scores during treatment; TX/POST = scores during treatment as well as immediately after treatment (less than 3 weeks after treatment ended); MAINT = maintenance (3 or more weeks after treatment ended).

Table 5  
*Quality Means, Standard Deviations, and Percentages for Writing Intervention Studies Based on Treatment Type*

Treatment	Total quality score		N	%										
	M	SD		Participant description	Participant selection	Physical description	DV(s) Quantified	DV(s) Reliable	Multiple baseline data points	Multiple interven. data points	Treatment description	Fidelity of treatment	Testing procedure description	Social validity
All studies	7.90	1.70	76	51.32	47.37	30.26	94.74	90.79	92.11	90.79	93.42	43.42	68.42	59.21
Strategy instruction plan/draft	8.87	1.19	25	68.00	52.00	12.00	96.00	92.00	96.00	96.00	100.00	84.00	84.00	80.00
Teaching grammar/usage	8.18	0.77	4	50.00	50.00	50.00	100.00	100.00	100.00	100.00	100.00	0.00	75.00	75.00
Goal setting for productivity	8.16	1.03	7	28.57	28.57	42.86	100.00	100.00	100.00	100.00	85.71	42.86	85.71	42.86
Strategy instruction editing	7.94	1.76	5	40.00	80.00	40.00	80.00	100.00	60.00	80.00	100.00	40.00	20.00	80.00
Word processing	7.36	2.10	5	60.00	80.00	40.00	100.00	80.00	80.00	100.00	100.00	0.00	60.00	40.00
Reinforcement	6.99	1.23	7	14.29	14.29	28.57	100.00	100.00	100.00	100.00	100.00	0.00	85.71	28.57
Prewriting activities	8.00	1.10	5	100.00	40.00	60.00	100.00	100.00	100.00	80.00	100.00	0.00	100.00	40.00
Sentence construction	6.68	1.93	5	60.00	40.00	20.00	100.00	60.00	100.00	60.00	80.00	40.00	20.00	60.00
Strategy instruction paragraph construction	6.40	0.94	5	0.00	40.00	40.00	100.00	100.00	60.00		60.00	20.00	40.00	40.00
Self-monitoring	7.43	2.53	8	50.00	50.00	37.50	75.00	62.50	100.00	100.00	87.50	50.00	50.00	50.00

Note. Maximum number of quality points = 11. N = number of studies; DV = dependent variable; Interv. = intervention.



Paz, 1999, 2001; Mason, Snyder, Sukhram, & Kedem, 2006), and they are grouped in this fashion in Tables 2 and 3.

Twenty-one of the 25 studies had a common outcome measure: elements. These structural elements included the basic parts of a composition that students generated ideas for when planning. In 18 of the 21 studies that assessed elements, maintenance effects (3 or more weeks after the termination of treatment) for elements were assessed. In addition four studies examined generalization of elements from one genre to another (i.e., story to personal narrative). Separate PNDs (mean, median, and range) were computed for each of these conditions.

There were 108 students (almost equally divided between boys and girls) in the 21 planning/drafting intervention studies that included a measure of elements (see Table 2). These youngsters were in Grades 2–8, and most of them were struggling writers (18 of the 21 studies). There were a variety of students with disabilities, but the most common group was children with learning disabilities (other students with disabilities included students with ADHD, emotional and behavioral disorders, mild language delay or speech/language impairment, and mild mental retardation). In one study (Albertson & Bilingsley, 1997), gifted students were the target of instruction (typical writers were the focus in the other two studies). Race/ethnicity was reported in 15 studies, and the majority of students were Caucasian (55%; 48/87), followed by African American (37%; 32/87), Arab (5%; 4/87), Hispanic (2%; 2/87), and Asian (1%; 1/87). When locale was reported ( $n = 16$  studies), suburban was the most common location (56%; 9/16), followed by rural (25%; 4/16) and urban settings (19%; 3/16). Researcher most commonly delivered instruction (81%; 17/21), and all of the studies employed multiple-baseline designs. It is also important to note that the 18 studies that assessed maintenance of elements had the same basic characteristics described above, but the four studies that assessed generalization of elements to a different genre only included struggling writers in Grades 2, 3, 5, and 6.

Teaching students a planning/drafting strategy had a large impact on increasing the number of basic genre elements in their writing, and this effect maintained over time (see Table 3). Median PND for treatment and maintenance was 100% in both cases. Mean PND was 96% (range = 100% to 67%) and 90% (range = 100% to 25%), respectively. Teaching these strategies also had a moderate impact on enhancing the generalization of elements from an instructed genre to an uninstructed one. Median PND was 86%, and mean PND was 85% (range = 100% to 67%).

Ten of the 25 planning/drafting strategy instructional studies assessed productivity (i.e., number of words written) during or immediately after instruction (mostly struggling writers in Grades 2–8 but also typical writers in Grades 6–8), with 7 of these studies assessing maintenance of productivity (focusing almost exclusively on struggling writers in Grades 2–8, with some typical writers in Grades 7 and 8). Moreover, 5 of the 25 studies assessed and graphed quality of students writing immediately following instruction (this involved students in Grades 2–4 and Grades 7 and 8 with a mixture of struggling and typical writers).

Teaching students a planning/drafting strategy had a large impact on productivity and quality during and/or immediately following instruction (see Table 3). Median PND for productivity and quality were 95% and 99%, respectively. Mean PND was 91% (range = 100% to 68%) and 99% (range = 100% to 97%), respectively. In addition, students generally maintained productiv-

ity gains, as median and mean PND were 100% and 86%, respectively (range = 100% to 33%).

We also examined the impact of strategy instruction for planning/drafting on writing genres. We computed separate PNDs for stories and expository writing (we combined persuasive text and expository essays for this second category). As can be seen in Table 4, this treatment had a large effect on both genres in terms of elements and productivity (all median and mean PNDs were 91% or greater). The only exceptions involved maintenance of elements for expository writing (mean PND = 89%) and maintenance of productivity for stories (mean PND = 79%).

The studies examining the teaching of strategies for planning/drafting were of high quality (on average, 82% of the quality indicators were met in all studies; see Table 5). The only quality indicator that was not met in at least 50% of the studies was description of the setting where instruction took place. Most importantly, experimental control was established in all but three of the studies that were used to compute summary PNDs for elements, productivity, or quality.

In summary, strategy instruction was effective in enhancing the number of elements, written output, and quality of students' writing, and the effects for elements and productivity were maintained over time. In addition, when transfer to a different genre was measured, positive effects were also obtained. For the most part, these findings generalize best to struggling writers in Grades 2–8 but are also valid for typical writers in Grades 4–8. Considerable confidence can be placed in these findings, as studies were of high quality and experimental control was established in almost all of them. The only concern involves failure for all but a few of the studies to adequately describe the setting where instruction took place. This narrows any generalization about conditions under which this treatment is effective.

### *Teaching Grammar/Usage*

Four studies evaluated the effectiveness of teaching grammar/usage (see Table 2). Teaching grammar/usage included peers directly teaching capitalization skills to classmates (Campbell, Brady, & Linehan, 1991) to teachers directly teaching adverbial phrases and possessives (Dowis, 1991; Dowis & Schloss, 1992) to how to correct capitalization, subject/verb agreements, conjunctions, incomplete sentences, and run-on sentences (Hermann, Semb, & Hopkins, 1976). All four studies included a similar outcome measure that focused on correct use of grammar. In all instances, the grammar measure assessed grammar skills students were taught during treatment.

There were 30 students in the four grammar/usage studies (see Table 2). The participants were all struggling writers in Grades 2, 5, and 6. In all but one of the studies (Hermann et al., 1976), the students were either learning disabled or classified as having a mild mental retardation (1 child). In the three studies that provided information on gender, boys and girls were represented almost equally.

In the two studies that provided information on race/ethnicity, most students were Caucasians (71%; 5/7) followed by African Americans (29%; 2/7). With the exception of the one study where peers delivered instruction (Campbell et al., 1991), teachers were the instructors. The effectiveness of grammar/usage instruction was tested via multiple-baseline and withdrawal designs.



Directly teaching grammar/usage had a moderate effect on improving grammar skills (see Table 3). The median PND was 84%, whereas the mean PND was 83% (range = 88% to 75%). On average, the four studies in this category met approximately 75% of the 11 quality indicators (see Table 5). None of the studies established treatment fidelity, but experimental control was established in three of the four investigations (see Table 2).

In summary, directly teaching grammar/usage had a positive impact on the grammar skills of struggling writers in Grades 2, 5, and 6. Studies were generally of high quality, but there was uncertainty that the treatment was delivered as intended in any of the studies.

### *Goal Setting for Productivity*

Seven studies were located that examined the impact of setting goals (see Table 2). Goal setting in these studies ranged from teachers encouraging students to exceed their previous publicly posted writing performance and receiving immediate feedback on their success (Van Houten, Morrison, Jarvis, & McDonald, 1974) to teachers setting a goal for how much students would write, with students monitoring their success in meeting the goal, and placing a star on a public chart if the goal was met (Wolfe, 1997). All of the studies included a common outcome measure: productivity. Productivity focused on amount of text produced and included the following measures: (a) number of words (Hopman & Glynn, 1989; Van Houten & McKillop, 1977; Van Houten et al., 1974; Wolfe, 1997), (b) number of words plus number of adjectives (Weygant, 1981), (c) number of writing lessons completed (Seabough & Schumaker, 1994), and (d) percentage of writing assignments completed (Kastelen, Nickel, & McLaughlin, 1984).

The seven goal setting studies included 129 students in Grades 2–5 and 8–12 (see Table 2). In the four studies that provided information on gender, the majority of participants were boys (85%, 17/20). Only one study reported any information on race/ethnicity (i.e., Weygant, 1981; all participants were Caucasian). The participants in four studies were youngsters with learning disabilities or other struggling writers, whereas youngsters in the remaining investigations represented a full range of abilities. Two of the studies were conducted in urban schools; the other five studies did not provide information on locale. In four studies, the instructor was the students' teacher, while researchers provided instruction in the other three studies. A variety of different designs were used to assess the effectiveness of goal setting for productivity, including changing criterion, multiple-baseline, and withdrawal designs.

Depending on which measure of central tendency, mean or median, was used to summarize the overall impact of the treatment, goal setting for productivity had a large to moderate effect (see Table 3). The median PND was 91%, whereas the mean PND was 79% (range = 100% to 26%).

On average, the seven studies in this category met approximately 75% of the 11 quality indicators (see Table 5). Less than one half of the studies, however, adequately described participants, selection of participants, or the physical setting for the treatment. A similar problem occurred in terms of establishing fidelity of treatment or collecting data on social validity. In addition, only three of the seven studies achieved experimental control (see Table 2).

In summary, goal setting had a large to moderate effect on increasing writing productivity of students in Grades 2–5 and 8–12 (this includes struggling and regular writers). Although studies were generally of high quality, just three of the seven studies established experimental control. Furthermore, failure to adequately describe participants, settings, and selection procedures in a majority of studies narrows generalizations about whom and under what conditions this treatment is effective. There was also uncertainty, at least in some studies, that the treatment was delivered as intended, and it was unclear whether teachers and students valued the treatment.

### *Strategy Instruction: Editing*

Five studies examined the effectiveness of strategy instruction for editing (see Table 2). The editing strategies used in these studies ranged from an editing strategy developed as part of the University of Kansas strategy curriculum (Beals, 1983; Schmidt, 1983; Schumaker et al., 1982; Scott, 1993) to an editing strategy designed to help students use spell checkers when editing (McNaughton, Hughes, & Ofiesh, 1997). All of the studies included a common outcome measure: errors corrected. Errors corrected ranged from spelling errors corrected (McNaughton et al., 1997) to the correction of a broad range of errors (e.g., spelling, usage, and punctuation).

There were 32 participants in the five editing strategy instruction intervention studies (see Table 2). These youngsters were in Grade 4 and Grades 8–12. Each of the studies included struggling writers with learning disabilities, but one study primarily involved students representing the full range of writing abilities (Beals, 1983). Most of the participants were boys (72%; 23/32). When race/ethnicity was reported (three studies), the majority of students were Caucasian (65%; 13/20), followed by African Americans (35%; 7/20). Studies occurred in a variety of locations (two urban, one suburban, one rural, and one not reported). Teachers taught the editing strategy in 60% of the studies, and effectiveness was evaluated in all studies using multiple-baseline designs.

Depending on which measure of central tendency, median or mean, was used to summarize the overall impact of the treatment, teaching an editing strategy had a large to moderate effect on correcting errors in writing (see Table 3). The median PND was 100%, whereas the mean PND was 84% (range = 100% to 50%).

The researchers met most of the quality indicators in studies examining editing strategy instruction (72%), but more than one half of the studies failed to adequately describe participants or the setting where instruction took place (see Table 5). Testing procedures were not adequately described in a majority of studies, nor was treatment fidelity. Experimental control was only achieved in two of the five studies (see Table 2).

In summary, teaching an editing strategy had a large to moderate impact on decreasing errors in the writing of students in Grade 4 and Grades 8–12 (struggling writers and regular writers). Although studies were generally of high quality, some caution must be applied in interpreting this finding, as only two studies (out of five) established experimental control. In addition, failure to adequately describe participants and settings in a majority of studies narrows generalizations about whom and under what conditions this treatment is effective. There was also some uncertainty in a



majority of the studies that the treatment was delivered as intended.

### *Word Processing*

Five studies evaluated the effectiveness of word processing (see Table 2). Word processing in these studies ranged from students using word processing to write versus handwriting (Brigman, 1994; Burnett, 1984; Christensen, 1993; Handley-More et al., 2003) to students using word processing with a reminder to use the editing features of word processing software (Jones, 1985). Four of the studies used a similar outcome measure: productivity. Productivity included (a) total number of words (Christensen, 1993; Jones, 1985), (b) total number of words used per session (Brigman, 1994), and (c) total number of words per minute (Handley-More et al., 2003).

There were 27 students in the four word processing studies that assessed writing productivity. All of the students were struggling writers in Grades 1, 4, and 5. In one study, these struggling writers were learning disabled (Handley-More et al., 2003), and in another investigation, they were gifted underachievers (Christensen, 1993). Two of every 3 students in these four studies were male, and the studies occurred in urban, suburban, and rural locations. Race/ethnicity was reported in two studies, and the majority of these participants were Caucasian (83%; 15/18) followed by African Americans (17%; 3/18). Teachers were the instructors in two studies, and researchers were the instructors in the other two investigations. The effectiveness of word processing on productivity was tested via alternating treatment, multiple-baseline, and withdrawal designs.

Word processing had a moderate effect on increasing students' productivity (see Table 3). The median PND was 75%, whereas the mean PND was 70% (range = 100% to 29%). On average, the four studies in this category met close to 70% of the 11 quality indicators (see Table 5). None of the studies, however, established treatment fidelity, and less than one half of them adequately described the setting where instruction took place or assessed social validity. Experimental control was established in one half of the studies.

In summary, word processing had a positive impact on increasing productivity of struggling writers in Grades 1, 4, and 5. Some caution must be exercised in interpreting this finding, as experimental control was only established in 50% of studies. Failure to adequately describe the setting in more than one half of studies narrows any generalization about the conditions under which this treatment is effective. There was also uncertainty in all of the studies whether the treatment was delivered as intended. It was unclear whether teachers or students valued this treatment.

### *Reinforcement*

Seven studies examined the effectiveness of using reinforcement to enhance writing performance (see Table 2). Reinforcement in these studies included teacher praise for using different parts of speech in stories (Blase-Maloney, Jacobson, & Hopkins, 1975), obtaining reinforcement through public posting of writing performance (Ballard & Glynn, 1975), and receiving group reinforcement for including a certain number of words, different words, and new words in compositions (Brigham, Graubard, & Stans, 1972).

Four of the studies included a common outcome measure: productivity. Productivity measures in these studies included number of sentences, number of words, number of different types of words, and number of different sentence beginnings produced during writing (Ballard & Glynn, 1975; Blase-Maloney & Hopkins, 1973; Blase-Maloney et al., 1975; Brigham et al., 1972).

There were 60 students in Grades 3–6 in the four studies examining reinforcement (see Table 2). The participants' disability status, gender, and race/ethnicity were not reported in any study. Only two of the studies provided information on type of writer (Brigham et al., 1972, worked with struggling writers, whereas students in Ballard and Glynn, 1975, represented the full range of writing abilities). Likewise, only two studies specified location (Ballard & Glynn, 1975, were in a rural location, while Blase-Maloney et al., 1975, were in an urban school). Teachers delivered the treatment, and multiple-baseline designs were used in all studies.

Reinforcement had a large effect on students' writing productivity (see Table 3); the median and mean PND were 100% and 96%, respectively (range = 100% to 84%). The overall quality of reinforcement studies was not strong (see Table 5). None of the studies established treatment fidelity, and less than one in three studies collected social validity information. In addition, adequate descriptions of participants, settings, or selection processes were rare. Finally, only two of the four reinforcement studies that assessed productivity established experimental control (see Table 2).

In summary, reinforcement had a large impact on increasing the writing productivity of students in Grades 3–6 (both struggling and regular writers). Caution must be applied in interpreting this finding, as studies were generally of low quality, and only two (out of four) studies that assessed productivity established experimental control. Failure to adequately describe participants, settings, and selection procedures in a majority of studies narrows generalizations about whom and under what conditions this treatment is effective. There was uncertainty that the treatment was delivered as intended, and it was unclear whether teachers and students valued it.

### *Prewriting Activities*

The effectiveness of five prewriting activities was examined in four publications. The prewriting activities included using a computer prewriting outline to generate and organize information (Channon, 2004), learning to use a graphic organizer for generating ideas prior to persuasive writing (Thanhouser, 1994), and learning to use a story web for generating ideas prior to writing (Zipprich, 1995). Four prewriting comparisons in three studies (Channon, 2004; Thanhouser, 1994; Zipprich, 1995) included a common measure: writing quality. Quality was assessed using a holistic measure (see Graham & Perin, 2007b).

There were 16 students in the studies that examined the impact of prewriting strategies on the quality of writing (see Table 2). Participants were in Grades 3–5 (Zipprich, 1995), 8 (Channon, 2004), and 12 (Thanhouser, 1994). All of the students in these studies were struggling writers with learning disabilities. The majority of the youngsters in these studies were male (94%; 15/16) and Hispanic (50%; 8/16). The remaining students were Caucasian (44%; 7/16) or African Americans (6%; 1/16). In all but one case, the intervention was delivered by the students' teacher. Two of the



three studies indicated geographic location, which were urban and suburban schools, respectively. The effectiveness of prewriting in the three studies was assessed via alternating treatment and multiple-baseline designs.

The median and mean PND for prewriting were similar: 55% and 52%, respectively (range = 84% to 13%). Thus, prewriting had a small effect on improving writing quality.

The researchers met most of the quality indicators in studies examining prewriting activities (73%), but less than one half of the studies adequately described participant selection, established treatment fidelity (this was not done in any study), or assessed social validity (see Table 5). Experimental control was only achieved in one of the four comparisons that assessed writing quality (see Table 2).

In summary, prewriting activities had a small impact on improving the quality of writing produced by struggling writers in Grades 3–5, 8, and 12. Although studies were generally of high quality, caution must be applied in interpreting this finding, as only one comparison (out of four) established experimental control. In addition, the value of the treatment to teachers and students was unclear, and there was uncertainty that the treatment was delivered as intended.

### *Sentence Construction*

Five studies examined the effectiveness of teaching sentence construction skills (see Table 2). Sentence construction in these studies included teaching strategies for writing different types of sentences (Beals, 1983; Eads, 1991; Johnson, 2005; Schmidt, Deshler, Schumaker, & Alley, 1988; these strategies were developed as part of the University of Kansas strategy curriculum) and teaching students how to combine simpler sentences into more complex ones (First, 1994). All of the studies included a common measure: complete sentences. Complete sentences were measured by calculating the percentage of complete sentences independently written by students.

There were 64 participants in the five sentence construction studies (see Table 2). These youngsters were in Grades 6–8 and 10–12, and the studies included more boys than girls (66%; 42/64). In all studies, there were students with disabilities (mostly learning), but two studies focused mainly on students representing the full range of writing ability (Beals, 1983; Johnson, 2005). Four studies provided information on race/ethnicity; the majority of students were Hispanic (47%; 26/55), followed by Caucasian (44%; 24/55), African American (7%; 4/55), and Hispanic and African American (2%, 1/55). In the three studies that specified locale, they were evenly split between urban, suburban, and rural locations. In all cases, instruction was delivered by teachers, and multiple-baseline designs were used to test the effectiveness of teaching sentence skills.

Teaching sentence construction was an effective treatment (see Table 3). The median PND was 83%, whereas the mean PND was 86% (range = 100% to 78%). However, none of the studies established experimental control (see Table 2), and the overall quality of studies was poor (on average, each study only met 61% of the 11 quality indicators; see Table 5). More than one half of the studies did not meet the following criteria: participant selection, physical description of setting, fidelity of treatment, and testing procedure description.

In summary, sentence construction was an effective practice in increasing the percentage of complete sentences produced by students in Grades 6–8 and 10–12 (for struggling and regular writers). Caution must be exercised in interpreting this finding, as the research in this area was particularly weak.

### *Strategy Instruction: Paragraph Construction*

Five studies examined the effectiveness of teaching students strategies for constructing paragraphs. These five studies were included in four publications (Dowell, Storey, & Gleason, 1994; Moran, Schumaker, & Vetter, 1981; Sonntag & McLaughlin, 1984; Wallace & Bott, 1989). Paragraph construction strategies ranged from teaching students strategies for organizing and/or writing expository paragraphs (Moran et al., 1981; Sonntag & McLaughlin, 1984; Wallace & Bott, 1989) to teaching students how to write descriptive paragraphs (Dowell et al., 1994). A conceptually similar outcome measure, writing elements, was used in four of the examinations. Writing elements for paragraphs involved determining whether basic parts of a paragraph (e.g., topic sentence and concluding sentence) were evident and/or correctly used.

There were 18 students in Grades 8–9 in the four strategy instruction for paragraph studies where elements were assessed (see Table 2). Two thirds of these students were struggling writers with learning disabilities, whereas the remaining students represented the full range of writing ability. In the three studies where gender was reported, 67% of students were boys. No information on race/ethnicity was provided in any study. Two of the studies took place in a suburban location, with one study conducted in an urban school (no location was given for the fourth study). In all instances but one study (Wallace & Bott, 1989), teachers taught the strategies. Only multiple-baseline designs were used to test the effectiveness of this treatment.

Teaching strategies for writing paragraphs had a large and positive impact on the schematic structure (i.e., elements) of students' paragraphs (see Table 3). The median PND was 100%, whereas the mean PND was 97% (range = 100% to 89%). However, studies were of poor quality (on average, only 58% of the 11 quality indicators were met; see Table 5). More than one half of the studies did not meet the following criteria: participant description, participant selection, physical description of treatment setting, treatment fidelity, testing procedure description, or social validity. None of the studies established experimental control.

In summary, teaching a paragraph strategy had a large effect on the paragraph writing of students in Grades 8 and 9 (struggling and regular writers). Considerable caution must be exercised in interpreting this finding, as none of the studies established experimental control. Moreover, failure to adequately describe participants, selection procedures, and settings in a majority of studies narrows generalizations about whom and under what conditions this treatment is effective. It was unclear whether the treatment was delivered as intended in the majority of studies, nor was it clear that teachers or students valued this treatment.

### *Self-Monitoring*

Eight studies examined the effects of self-monitoring procedures (see Table 2). Self-monitoring procedures in these studies



included monitoring and graphing: the number of words and sentences written (Goddard, 1998), on-task behavior (Harris, Graham, Reid, McElroy, & Hamby, 1994; Moran, 2004; Rumsey & Ballard, 1985), or on-task behavior plus written productivity (Wolfe, 1997). Seven studies included a similar measure: productivity. Productivity included (a) total number of sentences (Goddard, 1998), (b) total number of words (Goddard, 1998; Harris et al., 1994; Jackson, 1994; Moran, 2004; Rumsey & Ballard, 1985; Wolfe, 1997), and (c) total number of written items completed (Shimabukuro, Prater, Jenkins, & Edelen-Smith, 1999).

There were 35 students in the seven self-monitoring studies (Grades 2, 3, 5–7) that assessed productivity (see Table 2). All of these students were struggling writers, and the majority of youngsters (57%; 15/35) were classified as having learning disabilities or ADHD. Most of the students were boys (80%; 28/35). Race/ethnicity was only reported in three studies, and location was provided in four of them (all but one of these studies were in suburban schools). Teachers delivered the treatment in the majority of the investigations (57%; 4/7). The effectiveness of self-monitoring was assessed through multiple-baseline and withdrawal designs.

Depending on the central tendency measure used, mean or median, self-monitoring was either ineffective or had a small impact on writing productivity (see Table 3). The median PND was 43%, whereas the mean PND was 51% (range = 93% to 23%).

On average, the self-monitoring studies met two thirds of the 11 quality indicators (see Table 5). Six quality indicators were met by just one half or less of the studies. These included participant description, participant selection, physical description of setting, treatment fidelity, description of testing procedures, and social validity. In addition, less than one half of the seven self-monitoring studies that assessed productivity established experimental control (see Table 2).

In summary, the effectiveness of self-monitoring was unclear, as the mean PND was in the small effect range, but the median PND indicated that the treatment was ineffective. Moreover, considerable caution must be exercised in interpreting any conclusions for this treatment, as experimental control was established in only 43% of studies. Failure to adequately describe participants, settings, and selection procedures in the majority of studies narrows generalizations about whom and under what conditions this treatment is effective. In addition, only one half of the studies established treatment fidelity, weakening the claim that the independent variable was responsible for changes in students' writing behaviors.

## Discussion

Writing is a critical skill in an advanced technological society. The future aspirations of students who do not master the writing process are at risk, as strong writing skills are needed to attend college and obtain more than menial employment (Graham & Perin, 2007b). An important ingredient in ensuring that students become skilled writers involves teachers' use of effective writing practices. Previous meta-analyses have examined true- and quasi-experimental writing intervention studies to identify such practices (Bangert-Drowns, 1993; Bangert-Drowns et al., 2004; Goldring et al., 2003; Graham, 2006a; Graham & Harris, 2003; Graham &

Olinghouse, in press; Graham & Perin, 2007a, 2007b; Hillocks, 1986). This meta-analysis extends these efforts to single subject design writing intervention studies, providing a broader base for making evidence-based recommendations for teaching writing.

## Caveats and Limitations

Before summarizing the primary findings from this review, it is important to explore the limitations of such an analysis. First, a meta-analysis, like the one conducted here, involves aggregating the findings from individual studies to draw a more general conclusion about a treatment. The value and breadth of a general conclusion depends on a variety of factors, such as who participates in the studies (grade and type of learner) and the quality of the investigations. For example, it would be inappropriate to draw a broad conclusion about a treatment for all students if the treatment was only tested with primary grade students. Consequently, our generalizations about the effectiveness of specific treatments are appropriately restricted to the grades and types of students tested. They are also constrained by study quality. Of course, more confidence can be placed in a conclusion drawn from this review if it is also supported by findings from other types of research (e.g., experimental and quasi-experimental research). Consequently, we also relate the findings from this review to prior reviews (e.g., Graham & Perin, 2007a, 2007b; Hillocks, 1986). In fact, we think that decisions concerning evidence-based teaching practices are most productive when they are based on all available evidence, including qualitative, correlational, experimental, and single subject research (Graham & Harris, in press).

Second, the effect size metric, PND, used in this study can only be calculated when data for each assessment point is available in the research report. Fortunately, we had only to eliminate two studies that did not provide the needed information (many more would have been eliminated if we had used the regression effect method). Nevertheless, it was not possible to compute PND for overall writing quality in most of the studies included in this review (this variable was often measured but not graphed). This variable has been the primary focus of previous meta-analyses of true- and quasi-experimental writing intervention research (e.g., Graham & Perin, 2007a; Hillocks, 1986). It is also important to note that PND and Cohen's *d* (the effect size of choice in the meta-analyses of group experimental studies) are not directly comparable.

Third, some instructional procedures have been the focus of more single subject design research than others. For example, strategy instruction for revising may be an effective procedure (see Appendix), but there are not enough single subject studies available to draw even a tentative conclusion about its impact. Moreover, there was only one treatment, strategy instruction for planning/drafting, that involved 10 or more single subject design tests of its effectiveness. A similar problem existed in the most current meta-analysis of true- and quasi-experimental design writing intervention studies (Graham & Perin, 2007a), as only four treatments (strategy instruction, word processing, processing writing approach, and grammar instruction) yielded 10 or more effect sizes. Less confidence can be placed in evidence-based recommendation based on a small number of studies. There is clearly a need for additional replication as well as the study of new treatments. Between this review and Graham and Perin's (2007a) meta-

analysis, only 213 investigations of writing interventions were located. It is obvious that the research base in writing is unacceptably thin. Simply put, federal agencies, such as the Institute of Educational Sciences and the National Institutes of Health, need to make writing a priority, making monies available for conducting more writing intervention research and preparing new researchers.

Fourth, one concern with meta-analysis involves comparability of outcome measures on which the effect sizes are based. We addressed this problem by only computing a summary PND (mean, median, and range) for a treatment when there were at least four or more studies that graphed a conceptually similar measure (i.e., productivity). It must be noted, however, that there were many instances where these conceptually similar measures were not exactly the same. For example, productivity typically involved number of words written but also included number of sentences. This introduces some unwanted noise into the machinery of meta-analysis.

### *What Instructional Practices Improve Students' Writing?*

We were able to calculate summary PNDs (mean, median, and range) for 10 writing treatments. For all but one treatment, the median and mean PND are all above the ineffective range (50% or below) established by Scruggs et al. (1986). The one exception was self-monitoring, where the mean PND was slightly above 50%, but the median was below this level. We do not offer a recommendation for this treatment, as its effectiveness is questionable and research quality in this area is generally poor.

In summarizing our findings, we included a recommendation, the median and mean PND for the treatment, grade range of students tested for each treatment, and the type of student to whom the recommendation best applied (typical students, struggling writers, or both). We did not analyze the data race/ethnicity, as these statistics were absent in almost one half of the studies. We also indicated when treatment findings need to be interpreted more cautiously due to poor study quality. The 10 recommendations were ordered according to two factors. Treatments that were based on the strongest research evidence were presented before treatments based on weaker evidence (this was determined by considering both the establishment of experimental control and the quality of studies). When the strength of the evidence was generally equivalent, then treatments with larger median PNDs were presented before ones with smaller PNDs. We privileged median over mean PND, as the median is less susceptible to the influence of outliers and does not necessitate the use of an equal interval scale (PND is based on percentage).

1. Teach students strategies for planning/drafting both narrative and expository text (treatment/posttreatment median PND for elements, productivity, and quality = 100%, 95%, and 99%, respectively; mean PND = 96%, 91%, and 99%, respectively). The effects of this treatment were not only maintained over time (3 weeks or longer; median PND = 100% for both elements and productivity; mean PND = 90% and 86%, respectively) but were generalized to untaught genres (median PND for elements = 86%; mean PND = 85%). Considerable confidence can be placed in the effectiveness of this treatment with struggling writers in Grades 2–8 and

typical writers in Grades 4–8. These findings also provided support for the SRSD model (Harris & Graham, 1996) of strategy instruction, as all of these studies used this model to teach the target strategies.

2. Directly teach grammar skills to struggling writers (median PND for grammar skills = 84%; mean PND = 83%). Confidence can be placed in the effectiveness of this treatment with students in Grades 2, 5, and 6.
3. Set clear and specific goals to increase students' writing productivity (median PND for productivity = 91%; mean PND = 79%). Confidence can be placed in the effectiveness of this treatment with struggling writers and typical students in Grades 2–5 and 8–12.
4. Teach students strategies for editing their compositions (median PND for reducing errors = 100%; mean PND = 84%). The confidence that can be placed in the effectiveness of this treatment with struggling and typical writers in Grade 4 and Grades 8–12 must be tempered somewhat due to concerns about the quality of the research.
5. Make it possible for students to use word processing as a primary tool for writing (median PND for productivity = 75%; mean PND = 70%). The confidence that can be placed in the effectiveness of this treatment with struggling writers in Grades 1, 4, and 5 must be tempered somewhat because of concerns about the quality of the research.
6. Reinforce students for their writing productivity (median PND for productivity = 100%; mean PND = 96%). The confidence that can be placed in the effectiveness of this treatment with struggling and typical writers in Grades 3–6 must be tempered somewhat because of concerns about the quality of the research.
7. Engage students in prewriting activities for gathering and organizing ideas in advance of writing (median PND for writing quality = 55%; mean PND = 52%). The confidence that can be placed in the effectiveness of this treatment with struggling writers in Grades 3–5, 8, and 12 must be tempered because of concerns about the lack of experimental control in the majority of studies.
8. Teach students how to form complex sentences (median PND for complete sentences produced = 83%; mean PND = 86%). The confidence that can be placed in the effectiveness of this treatment with struggling and typical writers in Grades 6–12 must be tempered somewhat because of concerns about the quality of the research and lack of experimental control.
9. Teach students strategies for writing different types of paragraphs (median PND for paragraph elements = 100%; mean PND = 97%). The confidence that can be placed in the effectiveness of this treatment with struggling and typical writers in Grades 8 and 9 must be



tempered considerably because of concerns about the quality of the research and lack of experimental control.

### *How the Findings From This Meta-Analysis Support, Extend, and Contradict Prior Findings*

Although it is not possible to compare directly the findings from this meta-analysis and previous ones examining true- and quasi-experimental writing intervention studies, the findings from this review support and extend findings from these previous analyses. Consistent with previous reviews involving both group experimental studies (Graham, 2006a; Graham & Harris, 2003; Graham & Perin, 2007a, 2007b) and single subject design investigations (Graham, 2006a; Graham & Harris, 2003), this review provided strong support for the effectiveness of teaching planning/drafting strategies to typical and struggling writers. It also provided further validation of the effectiveness of the SRSD model (Harris & Graham, 1996). Additional work needs to be done in this area, however, as such strategic instruction has mostly involved only stories and persuasive text, it has not covered all grade levels, and there is virtually no evidence on whether its effects are maintained past 6 months.

Consistent with previous reviews of experimental group studies (Bangert-Drowns, 1993; Goldring et al., 2003; Graham & Perin, 2007a, 2007b), word processing also had a positive impact on students' writing. This analysis extends these previous findings by demonstrating the effectiveness of word processing with struggling writers (this was not specifically examined in prior reviews). Likewise, we found that prewriting activities enhanced the quality of text produced by struggling writers. A previous meta-analysis of true- and quasi-experimental studies by Graham and Perin (2007a, 2007b) found a similar impact for more typical writers.

Replications and extensions between the findings from our review and previous meta-analyses involving group experimental studies were further found for goal setting and sentence construction. Graham and Perin (2007a, 2007b) indicated that establishing clear and specific writing goals improved the quality of text produced by typical and struggling writers. Our review of single subject design studies also supports the effectiveness of this treatment with typical and struggling writers but extends its impact to writing productivity. In addition, Hillocks (1986) as well as Graham and Perin (2007a, 2007b) reported that sentence combining was an effective practice for improving the quality of students' writing (only one of the studies reviewed involved struggling writers). We found that teaching sentence construction skills (including sentence combining) had a positive impact on the number of complete sentences produced by both typical and struggling writers (although most studies involved struggling writers).

Our review of single subject design writing intervention studies provided evidence on the effectiveness of several writing treatments that were not examined in prior reviews. This included reinforcing students' writing productivity, teaching strategies for editing text, teaching strategies for constructing paragraphs, and self-monitoring of writing productivity. With the exception of self-monitoring (which involved only struggling writers and produced questionable effects), these treatments were effective with typical and struggling writers. Noticeably missing from this and other reviews (e.g., Graham & Perin, 2007a, 2007b) were recommendations specifically on teaching strategies for revising content.

Unfortunately, there are very few single subject design studies (see Appendix) or experimental studies (see Graham & Perin, 2007a) that examined the effectiveness of such instruction. Clearly, additional research is needed in this area.

One area where our results were at odds with the finding of other reviews involved the teaching of grammar/usage. In their meta-analyses of true- and quasi-experimental studies, Hillocks (1986) and Graham and Perin (2007a, 2007b) reported that traditional grammar instruction did not enhance the quality of students' writing. In a narrative review of the research in this area, Andrews et al. (2006) also indicated that such instruction did not enhance students' grammar in text. In contrast, we found that grammar/usage instruction had a positive impact on the correct use of grammar in the text written by struggling writers. One explanation for this difference involves the type of writers studied. The reviews of group experimental studies focused almost exclusively on typical writers, whereas the participants in the single subject design studies were all weaker writers. It is also possible that the difference in findings was due to the methods used to teach grammar. In the single subject design studies, grammar/usage was primarily taught by the teacher modeling how to correctly apply the skill in writing, followed by students practicing applying the skill, with teacher assistance as needed. This was not typically done in the group experimental studies. In any event, additional research is needed to determine effective procedures for improving grammar in the writing of both typical and struggling writers.

### *Quality of Research*

This review provided some important insights into the strengths and weaknesses of current single subject design writing intervention research. When we looked at all studies together, we found that researchers consistently quantified their dependent variables, collected multiple baseline and intervention data points for writing, and adequately described their treatment. However, many researchers failed to adequately describe their participants, the processes used to select them, or the setting in which instruction took place, narrowing generalizations about to whom and under what conditions treatments are effective. In addition, evidence that a treatment was implemented as intended was provided in just 4 out of every 10 studies, raising uncertainty about treatment fidelity. It must be noted, however, that failure to collect such data, especially for older studies, does not mean that the independent variable was in fact compromised, as it has only become common to report treatment fidelity in the last 10 years or so (Graham & Perin, 2007a). Two other areas that are in need of some improvement involved adequately describing testing procedures (this did not occur in 1 out of every 3 studies) and collecting information on the social validity of a treatment (this did not occur in 4 out of every 10 studies).

Perhaps most importantly, many researchers failed to establish experimental control for one or more graphed variables in their study (this occurred 40% of the time for studies included in Table 2). There was also considerable variability across treatment categories in terms of the quality of studies and establishment of experimental control. This led us to clarify how much confidence could be placed in specific recommendations. Our analysis of the quality of individual studies indicated that there is considerable room for improvement in single subject design writing interven-



tion research, especially in terms of describing who participates and where instruction took place, documenting treatment fidelity, describing testing procedures, assessing social validity, and establishing experimental control. Researchers and those who review future research in this area (e.g., as journal reviewers and dissertation committee members) need to ensure that these problems are addressed.

### *Issues Involved in Implementing the Recommendations*

Implementing research-based treatments is a challenging task. Just because a treatment was effective in a research study does not guarantee that it will be effective in all other situations. There is rarely, if ever, a perfect match between the conditions under which the research was implemented and the conditions in which it is subsequently put to use by practitioners. Even when the match is good, the safest course of action is to monitor continually the effects of the treatment to gauge directly whether it is effective under the new conditions.

It is also important to note that we do not know what combination or how much of each of the recommended treatments in this review or other reviews (e.g., Graham & Perin, 2007a) is needed to maximize writing instruction. There is some preliminary evidence, however, that integrating some treatments can be beneficial (see Danoff, Harris, & Graham, 1993; Sadoski, Wilson, & Norton, 1997). Furthermore, the recommendations for teaching writing from this and other recent reviews (e.g., Graham & Perin, 2007a) are incomplete, as they do not address all aspects of writing (e.g., writing vocabulary, classroom-based assessment, parental participation, and use of new technologies, motivation).

A final issue in implementing evidence-based writing practices and recommendations revolves around the different organizational structures or formats for teaching writing that exist in schools. At the elementary level, regular classroom teachers, special education teachers, other specialists (e.g., reading specialists), and aides may all be involved in one or more aspects of writing instruction. At the secondary level, writing instruction may occur within the context of the language arts or English classroom, other content classrooms (such as history or biology), or with a learning specialist (such as a special education teacher). At either level, writing might be taught as a separate subject, in conjunction with content learning in some classes or subjects, or infused throughout the curriculum. The effectiveness of these various formats has neither been tested nor compared one to another (Graham & Perin, 2007a). Furthermore, it is not certain how well the evidence-based practices identified in this or other reviews would fare in these different formats. Before implementing one or more of these treatments identified in this review, careful analysis of the organizational structure or format within which it will be placed should be undertaken, with the aim of identifying factors that may facilitate or impede effectiveness.

### *Concluding Comments*

Meta-analysis provides a useful tool for drawing "important insight from what might otherwise be a confused and disparate literature" (Bangert-Drowns et al., 2004, p. 52). The writing intervention literature in general, and single subject design studies in this area, clearly fit this description, as they tested the effectiveness

of a wide range of interventions. We capitalized on the strengths of meta-analysis to conduct a comprehensive review of writing treatments assessed via single subject design. This was a productive strategy, as we were able to draw nine evidence-based recommendations for teaching writing.

### References

References marked with an asterisk indicate studies included in the meta-analysis.

- Achieve, Inc. (2005). *Rising to the challenge: Are high school graduates prepared for college and work?* Washington, DC: Author.
- \*Adkins, M. H. (2005). *Self-regulated strategy development and generalization instruction: Effects on story writing among second and third grade students with emotional and behavioral disorders*. Unpublished doctoral dissertation, University of Maryland.
- \*Albertson, L. R. (1998). *A cognitive-behavioral intervention study: Assessing the effects of instruction on story writing*. Unpublished doctoral dissertation, University of Washington.
- \*Albertson, L. R., & Billingsley, F. F. (1997, March). *Improving young writers' planning and reviewing skills while story-writing*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.
- Andrews, R., Torgerson, C., Beverton, S., Freeman, A., Locke, T., Low, G., Robinson, A., & Zhu, D. (2006). The effects of grammar teaching on writing development. *British Educational Research Journal*, 32, 39–55.
- \*Ballard, K. D., & Glynn, T. (1975). Behavioral self-management in story writing with elementary school children. *Journal of Applied Behavior Analysis*, 8, 387–398.
- Bangert-Drowns, R. (1993). The word processor as an instructional tool: A meta-analysis of word processing in writing instruction. *Review of Educational Research*, 63, 69–93.
- Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based Writing-to-Learn interventions on academic achievement: A meta-analysis. *Review of Educational Research*, 74, 29–58.
- \*Bauernschmidt, M. C. (1991). *Repeated writing to improve writing fluency for students with mild handicaps*. Unpublished doctoral dissertation, University of Wisconsin—Milwaukee.
- \*Beals, V. L. (1983). *The effects of large group instruction on the acquisition of specific learning strategies by learning disabled adolescents*. Unpublished doctoral dissertation, University of Kansas.
- \*Blase-Maloney, K. B., & Hopkins, B. L. (1973). The modification of sentence structure and its relationship to subjective judgments of creativity in writing. *Journal of Applied Behavior Analysis*, 6, 425–433.
- \*Blase-Maloney, K. B., Jacobson, D. R., & Hopkins, B. L. (1975). An analysis of the effects of lectures, requests, teach praise and free time on the creating writing behavior of third grade children. In E. Ramp & G. Semb (Eds.), *Behavior analysis* (pp. 244–260). Englewood Cliffs, NJ: Prentice-Hall.
- \*Bording, C., McLaughlin, T. R., & Williams, R. L. (1984). Effects of free time on grammar skills of adolescent handicapped students. *Journal of Educational Research*, 77, 312–318.
- \*Boyer, A. W. (1990). *Improving the expository paragraph writing of learning disabled elementary school students using small group strategies instruction and word processing*. Unpublished doctoral dissertation, University of Kentucky.
- \*Brigham, T. A., Graubard, P. S., & Stans, A. (1972). Analysis of the effects of sequential reinforcement contingencies on aspects of composition. *Journal of Applied Behavior Analysis*, 5, 421–429.
- \*Brigman, D. J. P. (1994). *The effects of writing with pencil and paper compared with writing with a computer and a word processing program on first grade students who have writing difficulty*. Unpublished doctoral dissertation, Johns Hopkins University.



- \*Burnett, J. H. (1984). *Word processing as a writing tool of an elementary school student*. Unpublished doctoral dissertation, University of Maryland.
- \*Campbell, B. J., Brady, M. P., & Linehan, S. (1991). Effects of peer-mediated instruction on the acquisition and generalization of written capitalization skills. *Journal of Learning Disabilities*, 24, 6–14.
- \*Campbell, J. A., & Willis, J. (1978). Modifying components of creative behavior in the natural environment. *Behavior Modification*, 2, 549–564.
- \*Channon, S. (2004). *Effects of writing support programs on the writing quality and attitudes of Hispanic junior high school students with learning disabilities*. Unpublished doctoral dissertation, Illinois State University.
- \*Christensen, E. E. (1993). *The effects of word processing on the creative writing of high achieving and low achieving gifted elementary students*. Unpublished doctoral dissertation, University of Washington.
- \*Cole, K. B. (1992). *Efficacy and generalization of instruction in sequential expository writing for students with learning disabilities*. Unpublished doctoral dissertation, Northern Illinois University.
- \*Danoff, B., Harris, K. R., & Graham, S. (1993). Incorporating strategy instruction within the writing process in the regular classroom: Effects on the writing of students with and without learning disabilities. *Journal of Reading Behavior*, 25, 295–322.
- \*De La Paz, S. (1999). Self-regulated strategy instruction in regular education settings: Improving outcomes for students with and without learning disabilities. *Learning Disabilities Research & Practice*, 14, 92–106.
- \*De La Paz, S. (2001). Teaching writing to students with attention deficit disorders and specific language impairment. *Journal of Educational Research*, 95, 37–47.
- \*De La Paz, S., & Graham, S. (1997). Strategy instruction in planning: Effects on the writing performance and behavior of students with learning difficulties. *Exceptional Children*, 63, 167–181.
- \*Dowell, H. A., Storey, K., & Gleason, M. (1994). A comparison of programs designed to improve the descriptive writing of students labeled learning disabled. *Developmental Disabilities Bulletin*, 22, 73–91.
- \*Dowis, C. L. (1991). *The effects of mini-lesson instruction on the writings of students with learning disabilities within the writing process using whole group instruction*. Unpublished doctoral dissertation, University of Missouri—Columbia.
- \*Dowis, C. L., & Schloss, P. (1992). The impact of mini-lessons on writing skills. *Remedial and Special Education*, 13, 34–42.
- \*Eads, J. R. (1991). *Classroom teacher mediated generalization of a sentence writing strategy to the regular classroom by nine middle school students with learning disabilities*. Unpublished educational specialist thesis, Northeast Missouri State University.
- \*First, C. G. (1994). *The effects of sentence combining on the written expression skills of students with serious emotional disturbances*. Unpublished doctoral dissertation, University of the Pacific.
- \*Germain, J. C. (2004). *Remediation of written expression deficits in an elementary school population*. Unpublished doctoral dissertation, University of Northern Colorado.
- \*Glomb, N., & West, R. P. (1990). Teaching behaviorally disordered adolescents to use self-management skills for improving the completeness, accuracy, and neatness of creative writing homework assignments. *Behavioral Disorders*, 15, 233–242.
- \*Goddard, Y. L. (1998). *Effects of self-monitoring and self-evaluation on the written language performance and on-task behavior of elementary students with learning disabilities*. Unpublished doctoral dissertation, Ohio State University.
- Goldring, A., Russell, M., & Cook, A. (2003). The effects of computers on student writing: A meta-analysis of studies from 1992–2002. *Journal of Technology, Learning, and Assessment*, 2, 1–51.
- Graham, S. (1999). Handwriting and spelling instruction for students with learning disabilities. *Learning Disabilities Quarterly*, 22, 78–98.
- Graham, S. (2006a). Strategy instruction and the teaching of writing: A meta-analysis. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 187–207). New York: Guilford Press.
- Graham, S. (2006b). Writing. In P. Alexander & P. Wiene (Eds.), *Handbook of educational psychology* (pp. 457–477). Mahwah, NJ: Erlbaum.
- Graham, S., & Harris, K. R. (2003). Students with learning disabilities and the process of writing: A meta-analysis of SRSD studies. In L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of research on learning disabilities* (pp. 383–402). New York: Guilford Press.
- Graham, S., & Harris, K. R. (in press). Evidence-based writing practices: Drawing recommendations from multiple sources. *British Journal of Educational Psychology* (monograph series).
- \*Graham, S., & MacArthur, C. (1988). Improving learning disabled students' skills at revising essays produced on a word processor: Self-instructional strategy training. *Journal of Special Education*, 22, 133–152.
- \*Graham, S., MacArthur, C., Schwartz, S., & Page-Voth, V. (1992). Improving the compositions of students with learning disabilities using a strategy involving product and process goal setting. *Exceptional Children*, 58, 322–334.
- Graham, S., & Olinghouse, N. (in press). Learning and teaching writing. In E. Anderman & L. Anderman (Eds.), *Psychology of classroom learning*. Farmington Hills, MI: Gale.
- Graham, S., & Perin, D. (2007a). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, 99, 445–476.
- Graham, S., & Perin, D. (2007b). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools*. New York: Carnegie Corporation of New York.
- \*Guzel-Ozmen, R. (2006). The effectiveness of modified cognitive strategy instruction in writing with mildly mentally retarded Turkish students. *Exceptional Children*, 72, 281–297.
- \*Handley-More, D., Deitz, J., Billingsley, F. F., & Coggins, T. E. (2003). Facilitating written work using computer word processing and word prediction. *American Journal of Occupational Therapy*, 57, 139–151.
- \*Harris, K. R., & Graham, S. (1985). Improving learning disabled students' composition skills: Self-control strategy training. *Learning Disability Quarterly*, 8, 27–36.
- Harris, K. R., & Graham, S. (1996). *Making the writing process work: Strategies for composition and self-regulation* (2nd ed.). Cambridge, MA: Brookline Books.
- \*Harris, K. R., Graham, S., Reid, R., McElroy, K., & Hamby, R. S. (1994). Self-monitoring of attention versus self-monitoring of performance: Replication and cross-task comparison studies. *Learning Disability Quarterly*, 17, 121–138.
- Hayes, J. R. (2000). A new framework for understanding cognition and affect in writing. In R. Indrisano & J. R. Squire (Eds.), *Perspectives on writing: Research, theory and practice* (pp. 6–44). Newark, DE: International Reading Association.
- \*Hermann, J. A., Semb, S., & Hopkins, B. L. (1976). Effects of formal "grammar" and "direct method" training on the number of errors in compositions written by sixth-graders. *Revista Mexicana de Analisis de la Conducta*, 2, 68–84.
- Hillocks, G. (1986). *Research on written composition: New directions for teaching*. Urbana, IL: National Council of Teachers of English.
- \*Hopman, M., & Glynn, T. (1989). The effect of correspondence training on the rate and quality of written expression of four low achieving boys. *Educational Psychology*, 9, 197–213.
- Horner, R., Carr, E., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, 71, 165–180.
- \*Jackson, H. I. (1994). *Performance feedback and generalization programming in writing instruction*. Unpublished doctoral dissertation, Syracuse University.



- \*Jerram, H., Glynn, T., & Tuck, B. (1988). Responding to the message: Providing a social context for children learning to write. *Educational Psychology*, 8, 31–40.
- \*Johnson, C. S. (2005). *Teaching sentence writing preskills to middle school students with mild to moderate disabilities*. Unpublished master's thesis, California State University, Fullerton.
- \*Jones, M. L. D. (1985). *Effects of a microcomputer word processor on the writing and editing of fourth grade reluctant writers*. Unpublished doctoral dissertation, University of Maryland.
- \*Kastelen, L., Nickel, M., & McLaughlin, T. F. (1984). A performance feedback system: Generalization of effects across tasks and time with eighth-grade English students. *Education and Treatment of Children*, 7, 141–155.
- Kennedy, C. H. (2002). The maintenance of behavior change as an indicator of social validity. *Behavior Modification*, 26, 594–604.
- \*Korducki, R. A. (2001). *An instructional program integrating strategies for composition and self-regulation: Effects on the English and Spanish language writing skills of bilingual Latino students with learning difficulties*. Unpublished doctoral dissertation, University of Wisconsin—Milwaukee.
- \*Kraetsch, G. A. (1981). The effects of oral instructions and training on the expansion of written language. *Learning Disability Quarterly*, 4, 82–90.
- \*Lane, K. L., Harris, K. R., Graham, S., Weisenbach, J. L., Brindle, M., & Morphy, P. (2008). The effects of self-regulated strategy development on the writing performance of second-grade students with behavioral and writing difficulties. *Journal of Special Education*, 41, 234–253.
- \*Li, D. (2000). *Effect of story mapping and story map questions on the story writing performance of students with learning disabilities*. Unpublished doctoral dissertation, Texas Tech University.
- \*Lienemann, T. O. (2006). *Improving the writing performance of students with attention-deficit/hyperactivity disorder*. Unpublished doctoral dissertation, University of Nebraska—Lincoln.
- \*Lienemann, T. O., Graham, S., Leader-Janssen, B., & Reid, R. (2006). Improving the writing performance of struggling writers in second grade. *Journal of Special Education*, 40, 66–78.
- Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- \*Martin, K. F., & Manno, C. (1995). Use of a check-off system to improve middle school students' story compositions. *Journal of Learning Disabilities*, 28, 139–149.
- \*Mason, L. H., & Shriner, J. G. (2008). Self-regulated strategy development instruction for writing an opinion essay: Effects for six students with emotional/behavior disorders. *Reading & Writing: An Interdisciplinary Journal*, 21, 71–93.
- \*Mason, L. H., Snyder, K. H., Sukhram, D. P., & Kedem, Y. (2006). TWA + PLANS strategies for expository reading and writing: Effects for nine fourth-grade students. *Exceptional Children*, 73, 69–89.
- \*McCurdy, M. (2002). *Examining the effects of a multi-component writing program in increasing three writing skills of secondary school students*. Unpublished doctoral dissertation, Mississippi State University.
- \*McGee, S. J. (1996). *The effects of goal setting and attributional feedback on self-efficacy for writing and writing achievement*. Unpublished doctoral dissertation, Florida International University.
- \*McNaughton, D., Hughes, C., & Ofiesh, N. (1997). Proofreading for students with learning disabilities: Integrating computer and strategy use. *Learning Disabilities Research & Practice*, 12, 16–28.
- \*Montague, M., & Leavell, A. G. (1994). Improving the narrative writing of students with learning disabilities. *Remedial and Special Education*, 15, 21–33.
- \*Moran, M. R., Schumaker, J. B., & Vetter, A. F. (1981). *Teaching a paragraph organization strategy to learning disabled adolescents* (Research Rep. No. 54). Lawrence, KS: Institute for Research in Learning Disabilities.
- \*Moran, S. A. (2004). *Self-monitoring of attention and self-monitoring of performance with second grade journal writing: A comparison of two techniques*. Unpublished doctoral dissertation, University of Maryland.
- National Commission on Writing. (2003, April). *The neglected "R": The need for a writing revolution*. Retrieved April 23, 2008, from [http://www.writingcommission.org/prod\\_downloads/writingcom/neglect-edr.pdf](http://www.writingcommission.org/prod_downloads/writingcom/neglect-edr.pdf)
- National Commission on Writing (2004, September). *Writing: A ticket to work . . . or a ticket out: A survey of business leaders*. Retrieved April 23, 2008, from [http://www.writingcommission.org/prod\\_downloads/writingcom/writing-ticket-to-work.pdf](http://www.writingcommission.org/prod_downloads/writingcom/writing-ticket-to-work.pdf)
- National Commission on Writing (2005, July). *Writing: A powerful message from state government*. Retrieved April 23, 2008, from [http://www.writingcommission.org/prod\\_downloads/writingcom/powerful-message-from-state.pdf](http://www.writingcommission.org/prod_downloads/writingcom/powerful-message-from-state.pdf)
- \*Nelson, J. R., Smith, D. J., & Dodd, J. M. (1992). The effects of teaching a summary skills strategy to students identified as learning disabled on their comprehension of science text. *Education and Treatment of Children*, 15, 228–243.
- \*Newstrom, J., McLaughlin, T. F., & Sweeney, W. J. (1999). The effects of contingency contracting to improve the mechanics of written language with a middle school student with behavior disorders. *Child & Family Behavior Therapy*, 21, 39–48.
- Parker, R., Hagan-Burke, S., & Vannest, K. (2007). Percentage of all non-overlapping data (PAND): An alternative to PND. *Journal of Special Education*, 40, 194–204.
- Persky, H. R., Daane, M. C., & Jin, Y. (2003). *The nation's report card: Writing 2002* (NCES 2003–529). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.
- \*Regan, K. S. (2005). *Promoting expressive writing among students with emotional and behavioral disturbance via the dialogue journal*. Unpublished doctoral dissertation, George Mason University.
- \*Reid, R., & Lienemann, T. O. (2006). Self-regulated strategy development for written expression with students with attention deficit/hyperactivity disorder. *Exceptional Children*, 73, 53–68.
- \*Reier, S. B. (1998). *The effects of qualitative corrective feedback on the writing skills of students with learning disabilities using electronic mail dialogue journals*. Unpublished doctoral dissertation, Utah State University.
- \*Rumsey, I., & Ballard, K. D. (1985). Teaching self-management strategies for independent story writing to children with classroom behavior difficulties. *Educational Psychology*, 5, 147–157.
- \*Saddler, B. (2006). Increasing story-writing ability through self-regulated strategy development: Effects on young writers with learning disabilities. *Learning Disability Quarterly*, 29, 291–305.
- \*Saddler, B., Moran, S., Graham, S., & Harris, K. R. (2004). Preventing writing difficulties: The effects of planning strategy instruction on the writing performance of struggling writers. *Exceptionality*, 12, 3–17.
- Sadoski, M., Willson, V., & Norton, D. (1997). The relative contribution of research-based composition activities to writing improvement in lower and middle grades. *Research in the Teaching of English*, 31, 120–150.
- \*Schmidt, J. L. (1983). *The effects of four generalization conditions on learning disabled adolescents written language performance in the regular classroom*. Unpublished doctoral dissertation, Kansas State University.
- \*Schmidt, J. L., Deshler, D. D., Schumaker, J. B., & Alley, G. R. (1988). Effects of generalization instruction on the written language performance of adolescents with learning disabilities in the mainstream classroom. *Reading, Writing, and Learning Disabilities*, 4, 291–309.
- \*Schumaker, J. B., Deshler, D. D., Alley, G. R., Warner, M. M., Clark, F. L., & Nolan, S. (1982). Error monitoring: A learning strategy for improving adolescent academic performance. In W. M. Cruickshank & J. W. Lerner (Eds.), *Coming of age: Vol. 3. The best of ACLD* (pp. 170–183). Syracuse, NY: Syracuse University Press.



- \*Scott, K. S. (1993). *Generalization of cognitive strategies by students with learning disabilities: An instructional model*. Unpublished doctoral dissertation, University of Georgia.
- Scruggs, T. E., & Mastropieri, M. A. (1998). Summarizing single-subject research: Issues and applications. *Behavior Modification*, 22, 221-242.
- Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single subject research: Methodology and validation. *Remedial and Special Education*, 8, 24-33.
- Scruggs, T. E., Mastropieri, M. A., Cook, S. B., & Escobar, C. (1986). Early intervention for children with conduct disorders: A quantitative synthesis of single-subject research. *Behavioral Disorders*, 11, 260-271.
- \*Seabaugh, G. O., & Schumaker, J. B. (1994). The effects of self-regulation training on the academic productivity of secondary students with learning problems. *Journal of Behavior Education*, 4, 109-133.
- \*Sexton, M., Harris, K. R., & Graham, S. (1998). Self-regulated strategy development and the writing process: Effects on essay writing and attributions. *Exceptional Children*, 64, 295-311.
- \*Shimabukuro, S. M., Prater, M. A., Jenkins, A., & Edelen-Smith, P. (1999). The effects of self-monitoring of academic performance on students with learning disabilities and ADD/ADHD. *Education and Treatment of Children*, 22, 397-414.
- Smyth, J. (1998). Written emotional expression: Effect sizes, outcome types, and moderating variables. *Journal of Consulting and Clinical Psychology*, 66, 174-184.
- \*Sonntag, C. M., & McLaughlin, T. F. (1984). The effects of training students in paragraph writing. *Education & Treatment of Children*, 7, 49-59.
- \*Stoddard, B., & MacArthur, C. A. (1993). A peer editor strategy: Guiding learning-disabled students in response and revision. *Research in the Teaching of English*, 27, 76-103.
- \*Thanhouser, S. P. (1994). *Function over form: The relative efficacy of self-instructional strategy training alone and with procedural facilitation for adolescents with learning disabilities*. Unpublished doctoral dissertation, Johns Hopkins University.
- \*Troia, G. A., Graham, S., & Harris, K. R. (1999). Teaching students with learning disabilities to mindfully plan when writing. *Exceptional Children*, 65, 235-252.
- \*Van Houten, R., & McKillop, C. (1977). An extension of the effects of the performance feedback system with secondary school students. *Psychology in the Schools*, 14, 480-484.
- \*Van Houten, R., Morrison, E., Jarvis, R., & McDonald, M. (1974). The effects of explicit timing and feedback on compositional response rate in elementary school children. *Journal of Applied Behavior Analysis*, 7, 547-555.
- \*Van Houten, R., & Nau, P. A. (1980). The effects of two types of peer comments on the use of paragraphing in written composition by elementary schoolchildren. *Child Behavior Therapy*, 2, 55-63.
- \*Wade, L. K. (1988). *An analysis of the effects of a peer feedback procedure on the writing behavior of sixth grade students*. Unpublished doctoral dissertation, University of Kansas.
- \*Walker, B., Shippen, M. E., Alberto, P., Houchins, D. E., & Cihak, D. F. (2005). *Learning Disabilities Research & Practice*, 20, 175-183.
- \*Wallace, G. W., & Bott, D. A. (1989). Statement-pic: A strategy to improve the paragraph writing skills of adolescents with learning disabilities. *Journal of Learning Disabilities*, 22, 541-553.
- \*Weygant, A. D. (1981). *The effects of specific instructions and a lesson on the written language expression of learning disabled elementary school children*. Unpublished doctoral dissertation, University of Virginia.
- \*Wolfe, L. H. (1997). *Effects of self-monitoring on the on-task behavior and written language performance of elementary students with learning disabilities*. Unpublished master's thesis, Ohio State University.
- \*Zipprich, M. A. (1995). Teaching web making as a guided planning tool to improve student narrative writing. *Remedial and Special Education*, 16, 3-15.

Received October 30, 2007

Revision received March 26, 2008

Accepted March 27, 2008 ■

## Appendix

## Writing Instruction Treatments That Included Three or Less Studies

Study	Design	N	Writer type	Grade (Age)	Disability	Race/ethnicity	Location	Inst.	Quality Score (0-11)	PND TX	PND POST	PND MAINT	PND GEN
Dialogue journal Regan (2005)	MBD	5	SW	6 (11-12)	ED	AA (n = 2) C (n = 2) H (n = 1)	NR	Teacher	9.67	30% (Production) 20% (Quality)	35% (Production) 0% (Quality)		21% (Production) 38% (Quality) Setting
Direct instruction of a broad array of skills Walker et al. (2005)	MBD	3	SW	(15-16)	LD	AA (n = 2) H (n = 1)	URBAN	RES	10.00	96% <sup>a</sup> (Vocab.)		100% (Vocab.)	
McGee (1996)	MBD	5	SW	3-5	GIFTED	NR	NR	Teacher	8.67	6% (Production) 18% (Vocab.) 0% (Organize)			
Direct teaching self-regulation strategies Glomb & West (1990)	MBD	2	SW	HS	EBD	NR	NR	RES	7.17			63% (Production)	
Feedback Jerram et al. (1988)	ABAB	24	FR	5 (9-10)	—	NR	SUB	Teacher	4.40	20% (Production)			
Reier (1998)	MBD	4	SW	10-12	LD	NR	NR	Teacher	9.67	53% (Grammar)	80% (Grammar)		
Wade (1988)	ABA	17	FR	6	—	NR	RURAL	Peers	7.20	94% Group (Grammar) 50% Individual <sup>a</sup> (Grammar)			
Van Houten & Nau (1980)	MBD	1	SW	4	—	NR	NR	Peers	7.60	100% (Grammar)			
Goal setting for grammar / sentence construction McCurdy (2002)	MBD	9	SW	9	ADHD (n = 1) EBD (n = 1) MMR (n = 3) LD (n = 1) SLI (n = 3)	NR	URBAN	Teacher	8.67	74% (Grammar)		38% (Grammar)	
McGee (1996)	MBD	5	SW	3 to 5	GIFTED	NR	NR	Teacher	8.67	30% (Production) 9% (Vocab.) 15% (Organize)			
Repeated writing Bauernschmidt (1991)	ABAB	3	NR	5 to 6 (10-12)	EBD and LD	NR	NR	RES	6.5	63% <sup>a</sup> (Production) 47% (Vocab) 27% (Grammar)			

(Appendix continues)



## Appendix (continued)

Study	Design	N	Writer type	Grade (Age)	Disability	Race/ethnicity	Location	Inst.	Quality Score (0-11)	PND TX	PND POST	PND MAINT	PND GEN
Strategy instruction (revising / editing)													
Graham & MacArthur (1988)	MBD	3	SW	5 to 6 (10-11)	LD	NR	SUB	RES	8.67		64% (Total Revisions) 100% <sup>a</sup> Meaning-changing Revisions	84% (Total Revisions) 100% (Meaning-changing Revisions)	67% (Total Revisions) 100% (Meaning-changing Revisions) Handwriting 100% (Revisions)
Stoddard & MacArthur (1993)	MBD	6	SW	7 to 8	LD	NR	SUB	Teacher	6.50		100% (Revisions) 100% (Meaning-changing Revisions)	100% (Revisions) 100% (Meaning-changing Revisions)	100% (Revisions) 83% (Meaning-changing Revisions) Handwriting
Strategy instruction (other)													
Boyer (1990)	MBD	3	SW	6 (12-13)	LD	NR	NR	RES	9.58		100% <sup>a</sup> (Elements)		
Cole (1992)	MBD	12	SW	3 to 5	LD	NR	SEMI-URBAN	RES	8.83	100% <sup>a</sup> (Quality) 100% <sup>a</sup> (Production)	100% (Quality) 100% (Production)		100% (Quality) 100% (Production) Setting
Li (2000)	MBD	4	SW	4 to 5 (9-11)	LD	NR	SUB	RES	9.67	88% <sup>a</sup> (Production) 13% (Vocab.)		75% (Production) 25% (Vocab.)	
Nelson et al. (1992)	MBD	5	SW	4, 6, 8 (9-13)	LD	AA (n = 3) C (n = 1) H (n = 1)	URBAN	Teacher	9.67	100% (Production)	100% (Production)		100% (Production) Reading
Verbal encouragement													
Kraetsch (1981)	ABAB	1	SW	(12)	LD	NR	NR	RES	7.50	100% <sup>a</sup> (Production)			
Word processing plus													
Handley-Moore (2003)	ATD	3	SW	4 to 5 (10-11)	LD	NR	NR	RES	5.50	9% (Production)			
Channon (2004)	ATD	7	SW	8 (14-15)	LD	H	URBAN	Teacher	8.83	56% (Quality)			

*Note.* Dashes indicate that data were not reported. ATD = alternating treatment design; ABA and ABAB = withdrawal design; MBD = multiple-baseline design; N = Number of participants on single subject graph(s); AVG = average; NR = not reported; SW = struggling writer; FR = full range; ADHD = attention-deficit/hyperactivity disorder; EBD = emotional and behavioral disorder; ED = emotionally disturbed; LD = Learning Disability; MMR = mild mental retardation; SLI = speech or language impairment; AA = African American; C = Caucasian; H = Hispanic; SUB = suburban; Instr. = Instructor; RES = researcher; PND = percentage of nonoverlapping data; TX = during treatment; POST = immediately after treatment (less than 3 weeks after treatment ended); MAINT = maintenance (3 or more weeks after treatment ended); GEN = generalization (generalization to other settings or genres); Elements = basic parts of a genre or type of writing; Grammar = included number of complete and complicated sentences, sentences with adjectives, the percentage of correct placement of indentations in paragraphs, percentage of correct words in a sequence, or a more general number of correct grammar skills or points for correct writing and errors found; Production = included number of words, ideas, and/or T-units, as well as number of assignments completed; Vocab = included number of different types of words.

<sup>a</sup> Experimental control established.

# Primary Grade Writing Instruction: A National Survey

Laura Cutler  
University of Maryland

Steve Graham  
Vanderbilt University

A random sample of primary grade teachers ( $N = 178$ ; 97% female) from across the United States was surveyed about their classroom instructional practices in writing. Most of the participating teachers (72%) took an eclectic approach to writing instruction, combining elements from the 2 most common methods for teaching writing: process writing and skills instruction. Although 90% of the teachers reported using most of the writing instructional practices that were included in the survey, there was considerable variability between teachers in how often they used specific practices. The study provides support for the following 7 recommendations for reforming primary grade writing instruction: (a) increase amount of time students spend writing; (b) increase time spent writing expository text; (c) provide better balance between time spent writing, learning writing strategies, and teaching writing skills; (d) place more emphasis on fostering students' motivation for writing; (e) develop stronger connections for writing between home and school; (f) make computers a more integral part of the writing program; and (g) improve professional development for writing instruction in teacher education programs.

**Keywords:** writing, primary grades, instruction, composition, instructional practices

Reading and mathematics have received considerable attention in recent efforts to improve schooling in the United States. For example, the No Child Left Behind Act of 2001 focused considerable attention and effort on improving students' skills in each of these domains. However, the second of the traditional three *Rs*, writing, was not emphasized in the No Child Left Behind Act of 2001 and has been virtually nonexistent in the school reform efforts in this nation (despite the efforts of groups such as the National Commission on Writing working to change this situation; see <http://www.writingcommission.org>).

Why is writing the absent *R* in the school reform movement? One thing is for certain: It is not because students are developing the writing skills they need to be successful. Take for instance the findings from the most recent National Assessment of Educational Progress (NAEP; Persky, Daane, & Jen, 2003). The writing of two thirds or more of the students tested in Grades 4, 8, and 12 was below grade-level proficiency. The results from the previous NAEP report yielded similar findings (Greenwald, Persky, Ambell, & Mazzeo, 1999). Just as importantly, many youngsters leave high school lacking the writing skills needed for success in college or the world of work. College instructors estimate that 50% of high school graduates are not prepared for college-level writing demands (Achieve, Inc., 2005), whereas businesses in the United States spend \$3.1 billion annually for writing remediation (National Commission on Writing, 2004). In 2003, the National Com-

mission on Writing bluntly concluded that the writing of students "is not what it should be" (p. 7).

Another possible reason for why writing is not at the center of the school reform movement is that writing is not an important skill in the 21st century. This is also not the case. Youngsters who do not learn to write well are at a considerable disadvantage. At school, their grades are likely to suffer, especially in classes where written tests and reports are the primary means for assessing progress (Graham, 2006b). Weaker writers are also less likely than their more skilled classmates to use writing to support and extend learning in content classrooms. Recent meta-analyses have shown that writing can enhance content learning (Bangert-Drowns, Hurley, & Wilkinson, 2004; Graham & Perin, 2007c). Their chances of attending college are reduced, as writing is increasingly being used to evaluate applicants' qualifications. At work, writing is a gateway for employment and promotion, especially in salaried positions (see reports by the National Commission on Writing, 2004, 2005). Employees in business as well as government must be able to create clearly written documents, memorandum, technical reports, and electronic messages. Finally, participation in civic life and the community at large has increasingly required the ability to write, especially as the use of e-mail and text messaging has become so widespread.

The importance of writing and concerns about the writing competence of students led the National Commission on Writing, which was established by the College Board (representing more than 4,300 colleges and schools), to issue a report in 2003 indicating that writing needs to be placed "squarely in the center of the school agenda" (p. 3). This call to action also included several recommendations regarding writing instructional practices: increase amount of writing students do within and outside of school, assess students' progress in writing, use technology to advance the learning and teaching of writing, and better prepare teachers to teach writing.

---

*Editor's Note.* Joanna Williams served as the action editor for this article.—KRH

---

Laura Cutler, Department of Special Education, University of Maryland; Steve Graham, Department of Special Education, Vanderbilt University.

Correspondence concerning this article should be addressed to Steve Graham, Peabody 328, 230 Appleton Place, Vanderbilt University, Nashville, TN 37203. E-mail: [steve.graham@vanderbilt.edu](mailto:steve.graham@vanderbilt.edu)



One barrier to evaluating and implementing these suggestions and other recent recommendations for improving the teaching of writing (see *Writing Next*; Graham & Perin, 2007a, 2007c) is that researchers currently have little data on what writing instruction looks like in schools. They do not have a good sense of how much students write or what they write. They also do not know how much time is devoted to writing instruction; what writing skills, processes, or knowledge are taught to students; what methods are used to teach writing; how or even if technology is part of the writing program; or whether teachers assess students' writing progress. Without such information, it is difficult to determine what needs to be done. For instance, calls to increase the amount of time students write provide a reasonable part of the remedy, if students write infrequently. However, the potential value of this solution is less certain if students already write frequently.

In this study, we examined the writing instructional practices of primary grade teachers. Instead of concentrating on how writing was taught in a single school, district, or state, we focused instead on the teaching of writing across the nation. Consequently, we randomly sampled teachers from across the United States and asked them to complete a questionnaire about writing and writing instruction in their classroom. This approach to the study of classroom practices is based on the assumption that teachers are aware of the elements of their teaching and are able to relate this knowledge to questions about their teaching practices, just as other professionals can relate what they do when queried about their actions (Diaper, 1989). There is evidence that teachers are able to do this, as previous studies using survey methodology to examine teachers' literacy practices are corroborated by findings from observational research (see for instance Bridge & Hiebert, 1985; DeFord, 1985; Pressley, Rankin, & Yokoi, 1996).

We decided to focus on primary grade writing instruction for several reasons. One, it is especially important that students get off to a good start in writing. There is a growing consensus that waiting until later grades to address literacy problems that have their origin at the primary level is not particularly successful (Slavin, Madden, & Karweit, 1989). Thus, reform efforts in writing are likely to be hampered if the primary grades are not included (Graham & Perin, 2007b). Two, while NAEP has collected some limited data on writing practices nationwide, it does not solicit information on the teaching of writing in the primary grade (Applebee & Langer, 2006). Insufficient data about writing practices at this level increase the probability of recommending solutions that do not fit the problem.

There is surprisingly little data on the instructional writing practices of primary grade teachers. Most of the available information examines the writing programs of exceptional teachers or schools (see Pressley, Gaskins, Solic, & Collins, 2006; Pressley, Mohan, Bogaert, & Fingeret, 2005; Pressley et al., 1996; Pressley, Raphael, Gallagher, & DiBella, 2004; Rankin-Erickson, & Pressley, 2000; Wharton-McDonald, Pressley, & Hampston, 1998; Wray, Medwell, Fox, & Poulson, 2000). Although the practices of more typical teachers have been studied, much of this research is dated (e.g., Bridge & Hiebert, 1985; Christenson, Thurlow, Yseldyke, & McVicar, 1989; Fisher & Hiebert, 1990). We located only three studies conducted during the 1990s and 2000s that involved more typical teachers. It is unlikely, however, that the findings from the first two studies reviewed below provide a representative picture of writing instruction nationwide. Stahl,

Pagnucco, and Suttles (1996) examined three classes in one school where the principal made it clear that he ran the school and he expected his teachers to emphasize writing skills instruction. They also examined three classes in a new school, where the principal hand picked teachers who followed a whole language philosophy to literacy instruction. These schools represent each end of the writing instruction continuum ranging from explicit teaching to natural learning approaches (Graham, 2000). In the second study, Bridge, Compton-Hall, and Cantrell (1997) examined how writing instruction in a school district in Kentucky was influenced by statewide legislation on literacy instruction. While some aspects of this school reform effort mirrored what occurred in other states, only the practices of a single school district were studied, limiting the overall generalizability of the findings.

A third study, however, did provide representative information on contemporary writing instruction in the primary grades. Graham, Harris, Fink-Chorzempa, and MacArthur (2003) sent a survey questionnaire to a random sample of primary grade teachers from throughout the United States. Seventy percent of teachers completed the survey, and responders and nonresponders were similar in terms of gender, grade, type of school (public versus private), school location (urban, suburban, and rural), school size, computers per pupil, and expenditures per pupil. Although the primary purpose of this study was to identify the types of instructional adaptations that teachers made for struggling writers, it also provided information on specific aspects of their writing program.

The teachers in the Graham et al. (2003) investigation indicated that their students spent slightly more than 35 min writing each day. They further indicated that they spent a little more than 1 hr a day teaching writing, with most of this time being devoted to teaching mechanics, grammar, and usage. These basic skills were typically taught several times a week or more, whereas writing processes (planning and revising) were most often taught weekly to several times a week. Most teachers also conducted minilessons on writing, retaught skills, modeled writing processes, and conferred with students about their writing at least weekly or more often. The use of invented spellings, student choice in selecting writing topics, and allowing children to work at their own pace on writing assignments were relatively common practices, as was students helping each other and sharing their writing with peers. Less common was student use of computers during the writing period, with this occurring monthly or less in 60% of classrooms. It must be noted that there was considerable variability in teachers' responses to many of the items on the questionnaires, especially ones that asked them to estimate how much actual time was devoted to teaching or student writing.

Although the Graham et al. (2003) study provided needed information on contemporary instructional practices in writing, it was incomplete. Teachers were only asked how frequently they engaged in activities and procedures that could easily be adapted. The current study extends the Graham et al. (2003) investigation by examining a much broader range of instructional practices and activities, and it examines whether teachers' reported application of these procedures is related to grade level. To select these writing procedures, we first examined studies from 1980 to the present where primary grade teachers were asked to report their writing practices or researchers observed how they taught writing (this included studies of exceptional teachers and schools). We supplemented this analysis by examining current books on how to teach



writing to young children to locate other activities and practices not queried or observed in the studies analyzed. This does not mean that we surveyed every possible writing procedure, but it does provide some assurance that we asked teachers about practices that have been applied in the past and ones that are considered important by experts on the teaching of writing.

The activities and practices included in our survey instrument reflect the most common theoretical views of writing and writing instruction. This included cognitive/motivational (e.g., Hayes, 1996) and sociocultural models of writing (e.g., Schultz and Fecho, 2000), as well as instructional models that emphasize the explicit teaching of skills and strategies (see Graham, 2006b) and those that emphasize process and communication (see Vaca & Rasinski, 1992). We were particularly interested in determining whether primary grade teachers' writing programs reflected a process approach to writing instruction (emphasis is placed on the act of composing and instruction is mostly provided through informal means when the need arises), a skills-based approach (emphasis placed on systematic instruction of basic writing skills), or a combination of the two approaches. We hypothesized that teachers would take an eclectic approach to writing instruction, applying instructional procedures that cut across these two common approaches. We based this prediction on a study by Graham, Harris, Fink-Chorzempa, and MacArthur (2002), who found that most primary grade teachers had multifaceted beliefs about writing instruction, embracing both systematic skill instruction and informal learning methods, while emphasizing that the process of writing is more important than the product.

A second purpose of this study was to examine typical writing practices in order to draw recommendations for improving writing instruction in the primary grades. This included drawing new recommendations from the description of classroom practices generated by the teachers, as well as examining the value of the following four existing recommendations made by the National Commission on Writing (2003), indicating in parentheses where we placed our emphasis: (a) students should write more (examining how much students write text that is paragraph length or longer), (b) technology needs to be a more integral part of writing instruction (examining computer use in the writing program), (c) students progress in writing needs to be monitored (examining monitoring of writing progress in the classroom), and (d) teachers need to be better prepared to teach writing (examining teachers' perceptions of the quality of their preservice preparation to teach writing).

## Method

### *Sampling Procedures*

A random sampling procedure, stratified by grade level, was used to identify 294 first- through third-grade teachers from the population of primary grade teachers in the United States. The names were selected from a comprehensive list of 560,320 primary grade teachers in more than 72,000 public and private schools compiled by Market Data Retrieval (MDR). At each grade level, 98 teachers were randomly sampled from all public and private school teachers in the MDR data base (no other variables, such as geographic region, were used as part of the sampling scheme). We decided to use the MDR database, as this was used in the previous

national survey conducted by Graham et al. (2003), making the findings from the two studies more comparable. In addition, MDR provides information (grade, type of school, location, size of school, and school expenditure per pupil for commercial materials) for each person sampled. This allowed us to determine whether responders and nonresponders differed on several variables.

A sample size of 294 teachers (with 174 teachers completing the survey) is adequate for a population of 560,320 primary grade teachers under the following conditions (cf. Dillman, 2000): (a) a  $\pm 5\%$  sampling error is considered tolerable, (b) expected variation in teacher responses is equal to .125/.875, (c) the statistical confidence level is set at 95%, and (d) a return rate of 60% is obtained (this is a conservative estimate, as we obtained a return rate of 70% in Graham et al., 2003). Expected variation in teacher response rate refers to proportion of the population expected to choose a specific response to a question. The most common item in this survey (41 items) asked teachers to respond to an 8-point Likert-type scale. As a result, we used these items to determine expected variation in teacher response rate when determining needed sample size. Since we did not know what proportion of the population would be expected to choose each of the 8 points, we assumed an equal likelihood for selecting a point (yielding the .125/.875 split).

### *Survey Instrument*

Teachers were asked to complete a questionnaire that provided information about themselves, the composition of their classrooms, their attitudes and perceptions about writing and writing instruction, and their writing practices. The survey can be obtained by contacting Steve Graham.

The first section of the survey asked teachers to provide descriptive information on their gender, ethnicity, educational level, years spent teaching, and current grade level, as well as the class size and characteristics of their students (ethnicity, writing abilities, special needs, and socioeconomic status as determined by free and reduced lunch). This section also included a question asking teachers to rate (i.e., *exceptional*, *very good*, *adequate*, *poor*, *inadequate*) the quality of preparation to teach writing they received in their teacher certification program.

The second section included a series of four questions (using a 6-point Likert-type scale, with anchors ranging from *strongly agree* to *strongly disagree*) assessing teachers' attitudes towards writing ("I like to write") and teaching this skill, as well as their perceptions of their effectiveness to manage the writing classroom and teach writing.

The third section asked teachers to specify how much time their students spend writing each week and how much time they spend teaching specific skills and processes (i.e., spelling, handwriting, grammar, planning, and revising). Teachers were also asked to estimate what percentage of their instructional time was devoted to whole group, small group, and individualized instruction. They were further asked if they used a commercial program to teach any aspect of writing (and to identify the program if they did) and to identify the description that best described their approach to teaching writing: process writing, traditional skills instruction, a combination of the two, or some other approach (which the respondent was asked to describe). A final question asked them to indicate which writing activities students would work on during the academic year. This included a list of 20 options (ranging from



worksheets to writing to inform). Teachers were encouraged to identify any additional writing activities their students would complete that were not directly queried on the survey.

The fourth section included 41 Likert-type items. Each item focused on a specific activity or instructional procedure, and teachers were asked to indicate how often the practice was applied using an 8-point scale. The scale, developed by Pressley et al. (1996), included the following markers: 1 = *never*, 2 = *several times a year*, 3 = *monthly*, 4 = *several times a month*, 5 = *weekly*, 6 = *several times a week*, 7 = *daily*, and 8 = *several times a day*. The only exception involved three items ("invented spellings," "select own writing assignments," and "work at their own pace") where the 8-point scale ranged from *never* to *half the time to always*. The higher the score, the more often an activity or procedure occurred.

Eleven of the items asked teachers to indicate how often teachers, students, or both engaged in specific activities that supported the development of students' writing products (coefficient alpha = .78). These items focused on teacher/student conferences, student/student conferences, advanced planning, graphic organizers, revising, peers helping each other, writing at own pace, invented spellings, writing prompts, dictation, and computers for writing.

Six items examined how frequently teachers directly taught the following basic writing skills (coefficient alpha = .84): handwriting, spelling, capitalization, punctuation, grammar, and sentence construction. Another four items assessed how frequently writing processes were directly taught (coefficient alpha = .85), including modeling of writing strategies as well as teaching text organization, planning strategies, and revising strategies. Three items assessed more general instructional procedures (coefficient alpha = .62), including minilessons, reteaching, and setting multiple goals for writing lessons.

Six items addressed motivational activities and procedures (coefficient alpha = .70), including student selection of writing topics, sharing writing with peers, publishing, and working independently at writing centers, as well as teachers modeling their love of writing and sharing their writing with students. Four items assessed the use of the following assessment practices (coefficient alpha = .75): rubrics, writing portfolios, teacher monitoring of students' writing progress, and students' self-monitoring monitoring. Another four items assessed strategies used by teachers to extend writing to the home environment (coefficient alpha = .81): writing homework, writing at home with parental help, parents serving as an audience for child's writing, and teacher/parent communication. Finally, three items asked how often writing extended or interacted with other areas of the curriculum (coefficient alpha = .83): using writing to support reading, reading to support writing, and writing in other content areas.

For Sections 3 and 4 of the survey assessing teachers' writing practices, a pool of possible items were developed by first examining studies from 1980 to present where primary grade teachers were asked to report their writing practices or researchers observed how they taught writing (i.e., Bridge et al., 1997; Bridge & Hiebert, 1985; Christenson et al., 1989; Fisher & Hiebert, 1990; Graham et al., 2003; Pressley et al., 1996; Rankin-Erickson & Pressley, 2000; Stahl et al., 1996; Wharton-McDonald et al., 1998; Wray et al., 2000). A table containing each practice and activity observed or assessed was created, indicating which studies each assessed each procedure. Next, we examined current books on

teaching writing to young children (i.e., Culham, 1995; Duke & Bennett-Armistead, 2003; Gillet & Beverly, 2001; Mariconda, 2001; Nelson, Bahr, & Van Meter, 2004; Schaefer, 2001) and added recommended activities and practices to the table that had not been examined by researchers. With a few exceptions, we developed items for each of the practices and activities identified. We did not, for example, ask teachers if their students engaged in authentic writing (from Wharton et al., 1998) or scaffolding (from Pressley et al., 1996), as this term is too vague to be understood by respondents.

The instrument was field tested prior to its use in this study. Four primary grade teachers provided feedback on an initial version of the scale in terms of the clarity of each item. Their feedback resulted in changes to the wording of individual items, as well as the addition of the anchor point *several times a month* to the Likert-scale that ranged from *never* to *several times a day*.

We have collected some limited data showing that teachers' responses about their instructional practices are consistent with observations of their practices during writing period. In Lane et al. (2008), we asked 14 second-grade teachers to complete the 41 eight-point Likert-type items described above. We then conducted two observations of their writing instructional practices. With a few exceptions (i.e., we saw students plan and revise less frequently than teachers reported in the survey), the observations confirmed that the teachers applied the practices they reported using frequently in their classroom.

Furthermore, Olinghouse (2008) asked 13 second- and fourth-grade teachers to complete the items that asked teachers to specify exactly how much time students write and how much time the teachers spend teaching skills and processes (e.g., spelling, revising, etc.). They then observed each teacher's classroom on one occasion to determine whether reported time and observed time were similar. The reported and observed times were not statistically different.

## Procedure

A cover letter, the survey instrument, and a stamped return envelope were mailed to each teacher during the month of January. The cover letter indicated that we were conducting a survey to gather information on the teaching of writing at the primary grade level. Teachers were asked to return materials in the next 2 weeks if possible. To encourage completion and return of the survey, we included a \$2 bill in the package as a "thank you."

## Results

### Analyses

We first present information on the participating teachers. This includes comparing responders and nonresponders on the five variables (e.g., grade, type of school, etc.) provided by MDR. Next, using data from the survey, we describe the teachers and their students. This is followed by data about teachers' perceptions of their preparation to teach writing, their effectiveness as a writing teacher, and their attitudes toward writing and writing instruction. Finally, we examine the participants' reported teaching practices.

All together, the survey included 46 Likert-type items (5 assessed participants' preparation to teach writing and their and

attitude toward writing and writing instruction, whereas 41 items examined how often teachers, students, or both engaged in specific instructional activities and practices). Six items also asked teachers to indicate how much time during a typical week was devoted to specific activities (e.g., student writing, teaching handwriting, etc.), and 3 items asked teacher to indicate what percentage of their instructional time involved specific instructional arrangements (whole group, small group, and individualized). For all 54 of these items, we present the mean, standard deviation, and confidence interval (CI). We also examine whether there was a statistically significant relationship between grade level and teachers' responses using analysis of variance. When a statistically significant difference for grade was obtained, we used the least significant difference test to conduct post hoc analyses.

There were 20 items that asked teachers to indicate if their students would engage in specific writing activities during the year, 1 item that asked teachers to indicate their approach to writing instruction (e.g., process approach, skills approach, etc.), and 1 item that asked if materials were used to teach writing. Sampling error for these items was  $\pm 7.3\%$ . Chi-square analysis was used to determine whether there was a statistically significant difference by grade for each of these categorical variables. When a statistically significant chi-square was obtained, it was followed by a series of chi-squares comparing one grade to another.

In order to control for Type I errors, we set the critical alpha value using Bonferroni correction ( $\alpha = .05/68$ ) and rejected the null hypothesis if the  $p$  value was less than .00074. Using these criteria, there were few differences between teachers at different grade levels. Thus, the data presented in tables is for all teachers and grade level effects are addressed in text only if a statistically significant difference was obtained.

It is important to note that the responses for 30% of the items analyzed via analysis of variance did not meet the assumptions of homogeneity of variance due to skewness or kurtosis, with levels of one or more of these indices exceeding the range of  $\pm 1.96$  Fisher coefficient. For any variable where this was the case, we also ran the analysis using a nonparametric procedure—the Kruskal-Wallis procedure. In all cases, the results of parametric and nonparametric procedures were identical. Thus, we only report the parametric results in this article.

### *Participating Teachers and Their Students*

Of the 294 randomly selected primary grade teachers, 61% ( $n = 178$ ) agreed to participate in the study. Demographic information for the 178 responders as well as the 116 nonresponders is presented in Table 1. Chi-square analyses revealed no statistically significant differences between responders and nonresponders in terms of type of school (public vs. private), grade, or location of the school ( $p = .21, .75, .07$ , respectively). Analyses of variance further indicated that there were no statistically significant differences between responders and nonresponders in terms of school size or annual expenditures for materials per pupil ( $p = .29$  and  $.14$ , respectively). Thus, responders did not differ from nonresponders on these five demographic variables, providing evidence that the responders were representative of the sample as a whole.

Similar to previous surveys with primary grade teachers (Graham et al., 2003), almost all of the teachers were women (97%). For the most part, they were evenly distributed across the three

Table 1  
*Characteristics of Responders and Nonresponders*

Variable	Responders		Nonresponders	
	<i>n</i>	%	<i>n</i>	%
Type of school				
Public	158	89	107	92
Private	20	11	9	8
Grade				
First	58	33	42	36
Second	57	32	41	35
Third	63	35	33	28
Location				
Urban	39	22	33	29
Suburban	69	39	53	46
Rural	67	38	29	25
Size of school				
<i>M</i>	386.56		420.14	
<i>SD</i>	241.75		243.89	
Expenditures per pupil				
<i>M</i>	150.48		137.94	
<i>SD</i>	65.79		64.95	

grades. Two out of every 10 teachers were located in urban schools, with the remaining teachers split almost equally between suburban and rural locations. There was considerable variability in the size of the schools that employed the participating teachers.

As a group, the teachers had taught for 17 years ( $SD = 10$  years). They were overwhelmingly White (91%), with 4% of the teachers identifying themselves as Black, 2% as Hispanic, and 2% as other. Seventeen percent of the teachers had completed a master's degree plus additional coursework, 24% had completed a master's degree, 42% had completed a bachelor's degree plus additional coursework, and 16% had completed a bachelor's degree. Almost all of the teachers (92%) had attended a teacher certification program.

The average size of the participating teachers' class was 19.96 students ( $SD = 5.14$ ), with approximately 31% of their students qualifying for free or reduced-cost lunch. One tenth of their students were receiving special-education services, and the teachers indicated that 15% of their students were above average writers, 67% were average, and 21% were below average. In addition, 75% of the teachers' students were White, 11% were Hispanic, 10% were Black, 2% were Asian, and 2% were other.

### *Preparation to Teach Writing, Perceptions of Effectiveness, and Attitudes Toward Writing and Writing Instruction*

Of the 92% of teachers who had received certification through a teacher education program, 28% indicated that their preparation to teach writing was either very good or outstanding, 42% indicated that their preparation was adequate, and 28% indicated that it was poor or inadequate (sampling error =  $\pm 5.8\%$  for this item). Interestingly, the teachers' evaluation of their preparation to teach writing was not significantly related to how long they had taught ( $p = .06$ ). As can be seen in Table 2, teachers moderately agreed that they were effective writing teachers and managed their writing class effectively. They also moderately agreed that they liked to write and teach writing.



Table 2  
*Primary Grade Teachers' Perceptions of Their Effectiveness in Teaching Writing and Their Attitudes About Writing*

Survey question	n	M (SD)	CI
I like to teach writing.	177	5.0 (1.2)	4.8–5.1
I effectively manage my classroom during writing instruction.	177	5.1 (0.94)	4.9–5.2
I like to write.	177	4.7 (1.2)	4.6–4.9
I am effective at teaching writing.	175	4.8 (0.97)	4.7–4.9

Note. Scores range from 1 to 6, with higher scores indicating greater agreement. CI = 95% confidence interval.

Teaching Practices

*Approach to writing instruction and use of commercial materials.* When teachers were asked which best described their approach to writing instruction, 72% indicated that they used a process approach combined with a traditional skills approach, 20% a process approach, and 6% a traditional skills approach. For the 4 teachers who responded “other” approach, two of them indicated that they used the 6 + 1 trait method (Spandel, 2005), another provided a detailed description of their program (it included both process and skill instruction), and the fourth teacher referred to an approach that we were unable to identify.

For the most part, writing instruction in the primary grades was a home grown product: 65% of teachers reported that they did not use a commercial program to teach writing, handwriting, spelling, or any other aspect of writing (sampling error = ±7.3%). The other 35% of the teachers listed 137 different programs they used. The most common programs (45%) were designed to teach either handwriting or spelling. The next most common programs listed (36%) were basal language arts programs. The remaining programs ranged from 6 + 1 traits to writers’ workshop to skills-based writing programs.

*Organizational structure.* Teachers indicated they spend 56% (SD = 27%; CI = 52%–60%) of their instructional time with whole groups. Twenty-three percent (SD = 20%; CI = 20%–26%) of instructional time was devoted to small group instruction, with 24% (SD = 20%; CI = 21%– 27%) involving individualized assistance. There was considerable variability in how teachers organized their class for writing instruction (see standard deviations above).

*Writing activities.* Table 3 provides information on 20 writing activities and the percentage of teachers who reported that their students would complete each activity during the school year. At least 50% or more of the teachers have students work on 12 or more different writing assignments during the year, with the most common assignments involving story writing, drawing a picture and writing something to go with it, writing letters, journal writing, completing worksheets, composing a personal narratives, responding in writing to material read, and writing poems.

For five of the writing activities, there was a statistically significant difference by grade: writing summaries,  $\chi^2(2, N = 161) = 48.12, p < .00074$ ; writing to inform,  $\chi^2(2, N = 161) = 27.61, p < .00074$ ; writing to persuade,  $\chi^2(2, N = 161) = 36.28, p < .00074$ ; alphabet books,  $\chi^2(2, N = 161) = 17.44, p < .00074$ ; and biographies,  $\chi^2(2, N = 161) = 18.38, p < .00074$ ). For writing in

response to material read, follow-up analyses revealed that students in Grade 3 (93%) participated in this activity more than students in Grade 2 (72%; coefficient phi = .36) and Grade 1 (74%; coefficient phi = .28). Likewise, students in Grade 3 (97%) were more likely to write summaries than students in Grade 2 (60%; coefficient phi = .45) or Grade 1 (34%; coefficient phi = .66). Furthermore, Grade 2 students were more likely to write summaries than Grade 1 students (coefficient phi = .26). Similar results were obtained for writing to inform, as Grade 3 students (85%) were more likely to do this activity than were Grade 2 (56%; coefficient phi = .31) or Grade 1 students (36%; coefficient phi = .50). Similarly, students in Grade 3 (67%) were more likely to write to persuade than were students in Grade 2 (30%; coefficient phi = .37) or Grade 1 (13%; coefficient phi = .55). In contrast, students in Grade 1 (55%) were more likely to compose an alphabet book than students in Grade 3 (17%; coefficient phi = .39). Finally, Grade 3 students (50%) were more likely to write biographies than were Grade 2 (26%; coefficient phi = .25) or Grade 1 students (13%; coefficient phi = .39).

In addition to indicating whether their students completed the 20 listed writing activities, teachers were asked to identify any additional writing activities that their students completed over the course of the school year. Twenty-three percent (n = 40) of the teachers listed additional activities. The additional writing activities ordered from most frequent to least frequent were content area writing in math and science (21.7%), invitations and thank you notes (10.8%), descriptive writing (10.8%), newspaper articles (8.7%), instructions (8.7%), compare and contrast writing (8.7%), research reports (4.1%), writing to a prompt (4.7%), cause and effect writing (3.1%), note taking (3.1%), reflection/opinion writing (3.1%), and writing sentences (3.1%). In addition, 13 writing activities were listed only one time by respondents (i.e., phone

Table 3  
*Types of Writing Activities Students Do During the Academic Year by Frequency*

Writing activity	% of teachers who responded yes
Stories	96.1
Drawing a picture and writing something to go with it	94.9
Writing letters to another person	88.8
Journal writing	86.5
Completing worksheets	86.0
Personal narratives	79.8
Writing in response to material read	78.1
Poems	75.3
Writing summaries	65.7
Lists	65.2
Book reports	62.4
Writing to inform	59.0
Books	48.3
Copying text	42.7
Writing to persuade	36.0
Alphabet books	33.7
Autobiographies	29.2
Biographies	28.1
Comic strips	16.9
Plays	15.7

Note. Sampling error ±7.3%

messages, riddles, alliteration, dialogue, story problems, onomatopoeia, prayers, recipes, personal recounts, labels, cartoons, graphic organizers, and e-mailing letters to pen pals in another state).

We further asked teachers to indicate how many minutes each week their students spend writing text that was paragraph length or longer. This included planning, drafting, revising, and editing text. There was considerable variability in amount of time reported, ranging from 0 min to 380 min ( $SD = 70.8$  min;  $CI = \pm 6.5$  min). As a result, we report the median response instead of the mean. The median amount of time students composed was 105 min per week (i.e., 21 min per day).

*Practices that support students' writing.* Table 4 presents means, standard deviations, 95% confidence intervals, and percentage of teachers who checked each anchor point for 38 of the 41 Likert-type items (8-point scale, ranging from 0 [*never*] to 7 [*several times a day*]) that assessed the use of specific writing instructional practices. For each category of items (i.e., support students' writing, teach basic writing skills, teach writing processes and strategies, etc.), items are arranged from most to least frequently occurring. Means, standard deviations, and 95% confidence intervals for the other three Likert-type items are presented in text. Although these items also employed an

Table 4  
*How Frequently Primary Grade Teachers Use Specific Writing Practices*

Writing practice	Never	Several times/year	Once/month	Several times/month	Once/week	Several times/week	Daily	Several times/day	<i>M</i> ( <i>SD</i> )	<i>CI</i>
<b>Support student writing</b>										
Graphic organizers ( $n = 176$ )	6%	3%	12%	11%	26%	17%	11%	15%	4.1 (1.9)	3.8–4.4
Teacher conferences ( $n = 173$ )	2%	11%	8%	13%	30%	23%	11%	2%	3.8 (1.6)	3.6–4.0
Planning ( $n = 176$ )	1%	10%	10%	18%	26%	25%	11%	1%	3.8 (1.5)	3.6–4.0
Writing prompts ( $n = 177$ )	1%	14%	10%	20%	24%	18%	13%	1%	3.6 (1.6)	3.4–3.8
Revising ( $n = 177$ )	2%	15%	9%	27%	21%	19%	8%	0%	3.4 (1.6)	3.2–3.6
Helping peers with writing ( $n = 174$ )	11%	14%	12%	18%	16%	13%	13%	3%	3.2 (2.0)	2.9–3.5
Peer conferences ( $n = 174$ )	12%	17%	10%	13%	22%	15%	9%	2%	3.1 (1.9)	2.8–3.3
Computers ( $n = 175$ )	42%	25%	7%	3%	8%	7%	6%	1%	1.6 (2.0)	1.3–1.9
Dictation ( $n = 171$ )	56%	15%	7%	6%	5%	6%	5%	0%	1.2 (1.8)	1.0–1.5
<b>Teach basic writing skills</b>										
Spelling skills ( $n = 177$ )	0%	0%	0%	0%	19%	25%	51%	5%	5.4 (.85)	5.3–5.5
Capitalization skills ( $n = 177$ )	0%	3%	2%	3%	15%	17%	48%	12%	5.3 (1.4)	5.1–5.5
Grammar skills ( $n = 177$ )	1%	1%	1%	5%	15%	21%	51%	6%	5.3 (1.1)	5.1–5.4
Punctuation skills ( $n = 177$ )	0%	2%	2%	5%	15%	18%	49%	10%	5.3 (1.3)	5.1–5.5
Handwriting skills ( $n = 175$ )	3%	7%	0%	9%	22%	22%	35%	2%	4.6 (1.6)	4.3–4.8
Sentence construction skills ( $n = 177$ )	0%	4%	6%	12%	22%	28%	25%	3%	4.5 (1.4)	4.3–4.7
<b>Teaching writing process</b>										
Model writing strategies ( $n = 178$ )	2%	7%	10%	16%	26%	22%	17%	1%	3.9 (1.5)	3.7–4.2
Text organization skills ( $n = 174$ )	1%	13%	10%	21%	25%	20%	10%	0%	3.5 (1.5)	3.3–3.8
Strategies for planning ( $n = 175$ )	2%	14%	13%	19%	26%	21%	7%	0%	3.4 (1.5)	3.2–3.6
Strategies for revising ( $n = 177$ )	2%	18%	13%	20%	28%	12%	7%	0%	3.2 (1.6)	3.0–3.4
<b>General instructional procedures</b>										
Mini lessons ( $n = 176$ )	0%	6%	5%	9%	18%	30%	27%	6%	4.6 (1.5)	4.4–4.9
Multigoal lessons ( $n = 173$ )	0%	4%	13%	9%	26%	18%	15%	17%	4.4 (1.7)	4.2–4.7
Reteach skills ( $n = 174$ )	0%	13%	10%	29%	21%	16%	12%	1%	3.5 (1.5)	3.3–3.7
<b>Promoting motivation</b>										
Share writing with peer ( $n = 176$ )	0%	11%	11%	15%	34%	21%	8%	1%	3.7 (1.4)	3.5–3.9
Model enjoyment or love of writing ( $n = 174$ )	6%	17%	10%	15%	18%	16%	17%	2%	3.4 (1.9)	3.2–3.7
Publishing ( $n = 177$ )	3%	27%	23%	24%	17%	6%	1%	0%	2.4 (1.3)	2.3–2.7
Teacher reads own writing ( $n = 176$ )	19%	26%	9%	17%	11%	7%	10%	0%	2.4 (2.0)	2.1–2.7
Writing enters ( $n = 173$ )	35%	13%	6%	13%	16%	8%	9%	0%	2.2 (2.1)	1.9–2.5
<b>Assessment</b>										
Teacher monitor writing progress ( $n = 172$ )	1%	8%	8%	15%	27%	15%	26%	1%	4.1 (1.6)	3.9–4.4
Student monitor writing progress ( $n = 176$ )	5%	14%	9%	10%	17%	13%	30%	3%	3.9 (2.0)	3.7–4.2
Writing portfolios ( $n = 174$ )	32%	21%	10%	12%	10%	4%	10%	1%	2.0 (2.0)	1.7–2.3
Student use of rubrics ( $n = 175$ )	32%	20%	13%	18%	8%	6%	5%	0%	1.9 (1.8)	1.6–2.2
<b>Home environment</b>										
Writing homework ( $n = 175$ )	23%	26%	7%	13%	17%	5%	8%	1%	2.2 (2.0)	1.9–2.5
Students write at home with parental help ( $n = 172$ )	34%	28%	9%	6%	16%	3%	4%	0%	1.7 (1.8)	1.4–1.9
Parents listen to students' writing ( $n = 172$ )	29%	36%	8%	10%	12%	3%	2%	0%	1.6 (1.6)	1.3–1.8
Communicate with parents about students' writing progress ( $n = 175$ )	7%	65%	13%	6%	8%	1%	1%	0%	1.5 (1.1)	1.3–1.6
<b>Extend Writing to Content Areas</b>										
Cross-curriculum writing ( $n = 176$ )	2%	17%	9%	18%	23%	19%	11%	3%	3.6 (1.7)	3.3–3.8
Writing to support reading ( $n = 175$ )	3%	14%	9%	18%	18%	27%	10%	2%	3.6 (1.7)	3.4–3.7
Reading to support writing ( $n = 169$ )	8%	24%	10%	22%	14%	15%	7%	1%	2.9 (1.8)	2.6–3.1

Note. CI = 95% confidence interval.



8-point scale, the scale ranged from *never* (score of 0) to *always* (score of 8).

Two of the practices for supporting students' writing were assessed via the 8-point scale that ranged from *never* to *always*. One of these practices, encouraging students to use invented spellings, was applied frequently ( $M = 5.3$ ;  $SD = 1.9$ ;  $CI = 4.9-5.5$ ). Eighty four percent of teachers reported that they encouraged the use of invented spellings at least half of the time or more, with 37% indicating they always encouraged the use of this practice. Although students' selection of their own writing topics was less common ( $M = 3.6$ ;  $SD = 1.4$ ;  $CI = 3.4-3.8$ ), 63% of teachers reported applying this practice at least half of the time (only 3% always allowed students to chose their own writing topics).

As can be seen in Table 4, the majority of teachers reported that they used the following procedures to support students' writing at least weekly or more often: graphic organizers, writing prompts, teacher conferences with students about their writing, and planning in advance of writing. Several times a month or more often, the majority of teachers indicated they had students revise their written products. There was considerable variability in how often students helped each other with their writing or conferenced with their peers about what they wrote. Depending upon the teacher, these practices typically occurred somewhere between several times a month and daily. Much less frequently applied was the use of computers or dictation to support students' writing. Forty two percent and 56% of teachers, respectively, reported that they never used these practices. Only about one third of the participating teachers used computers or dictation at least once a month or more often.

There was a statistically significant grade-level difference for one of the writing support practices: using graphic organizers,  $F(2, 157) = 14.75$ ,  $MSE = 45.65$ ,  $p < .00076$ . Follow-up analyses revealed that Grade 3 teachers reported using graphic organizers more often than Grade 1 teachers ( $p < .001$ ; effect size [ES] = .90; Grade 1:  $M = 3.50$ ,  $SD = 1.6$ ; Grade 3:  $M = 5.2$ ,  $SD = 1.8$ ).

**Teaching basic writing skills.** Teachers reported that they frequently taught basic writing skills (see Table 4). Spelling, grammar, capitalization, and punctuation skills were taught by a majority of the teachers daily, with at least three quarters of them teaching these skills at least several times a week. The majority of teachers also reported teaching handwriting and sentence construction skills several times a week to daily, with three quarters of them teaching these skills at least weekly or more often. It is interesting to note that it was extremely rare for a teacher to report that these skills were never taught.

For one of the basic writing skills, handwriting, there was a statistically significant grade level difference,  $F(2, 155) = 9.78$ ,  $MSE = 24.05$ ,  $p < .00076$ . Post hoc analysis revealed that Grade 1 ( $p < .001$ ; ES = .76) and Grade 3 teachers ( $p < .001$ ; ES = .66) reported teaching handwriting more often than Grade 2 teachers (Grade 1:  $M = 5.0$ ,  $SD = 1.6$ ; Grade 2:  $M = 3.7$ ,  $SD = 1.8$ ; Grade 3:  $M = 4.8$ ,  $SD = 1.2$ ).

When we asked teachers to specify how much time a week they spent teaching spelling, grammar/usage, and handwriting, the patterns were similar. The typical teacher reported spending 74 min a week teaching spelling ( $SD = 61.6$ ;  $CI = 65.1-83.7$ ;  $Mdn = 60.0$ ) and 80 min a week teaching grammar/usage ( $SD = 76.7$ ;  $CI = 68.4-91.5$ ;  $Mdn = 60.0$ ), but only 46 min a week teaching handwriting ( $SD = 37.2$ ;  $CI = 40.8-52.1$ ;  $Mdn = 30.0$ ). Again,

virtually all of the teachers devoted at least some time each week to teaching these skills.

**Teaching writing processes and strategies.** The majority of teachers reported modeling how to apply writing strategies as well as teaching students about text organization and planning strategies at least once a week (see Table 4), with such instruction most commonly occurring once a week to several times a week. Revising strategies were taught by most teachers at least several times a month, with this mostly occurring on a weekly or greater basis. Only 1 in 50 teachers reported that they never taught writing strategies.

Consistent with these findings are the amount of time teachers indicated that they spent teaching planning and revising strategies each week (we report means and medians due to the large standard deviations). The typical teacher reported spending 38 min a week teaching planning strategies ( $SD = 28.3$ ;  $CI = 32.9-43.9$ ;  $Mdn = 30.0$ ) and 33 min a week teaching revising strategies ( $SD = 35.0$ ;  $CI = 28.3-37.8$ ;  $Mdn = 30.0$ ).

**Instructional procedures for teaching skills and strategies.** The majority of teachers reported that they conducted minilessons several times a week or more to teach needed skills and writing processes (see Table 4), with 80% of teachers doing this at least weekly. The reteaching of skills and strategies previously taught, however, occurred less frequently, as the majority of teachers reported doing this somewhere between several times a month to several times a week. There was considerable variability in how often teachers reported designing writing instruction with multiple goals, but the majority of teachers did this at least weekly.

**Motivational techniques.** Over 80% of the teachers reported that their students were allowed to work at their own pace on writing assignments at least half of the time or more ( $M = 4.8$ ;  $SD = 1.6$ ;  $CI = 4.5-5.0$ ), with approximately 1 in 5 teachers indicating that this always occurred. As can be seen in Table 4, a majority of the teachers indicated that students shared their writing with their peers at least weekly (no teacher responded "never" to this question). Although there was considerable variability in how often teachers modeled their enjoyment or love of writing, the majority of teachers also did this at least weekly. It was less common for teachers to share their own writing with students, have children publish their work, or work independently at a writing center. The majority of teachers did this monthly or less, and a sizable proportion of teachers did not use writing centers (35%) or share their writing with students (19%).

**Assessing writing.** The majority of teachers reported that they monitored their students' writing to make decisions about writing instruction at least weekly or more often (see Table 4). They also encouraged their students to monitor their own writing progress at a similar rate. In contrast, a sizable proportion of teachers reported that they did not have their students build writing portfolios (32%) or use rubrics to assess their writing (32%). When either of these practices were applied, they were more likely to occur several times a year to monthly.

There was a statistically significant grade-level difference in how often teachers reported monitoring students' writing progress,  $F(2, 154) = 9.84$ ,  $MSE = 22.10$ ,  $p < .00076$ .

Post hoc analyses revealed that Grade 1 ( $p < .001$ ; ES = .75) and Grade 3 teachers ( $p < .001$ ; ES = .68) reported that they were more likely to monitor the writing of their students than Grade 2



teachers (Grade 1:  $M = 4.5$ ,  $SD = 1.6$ ; Grade 2:  $M = 3.3$ ,  $SD = 1.6$ ; Grade 3:  $M = 4.4$ ,  $SD = 1.2$ ).

*Extending writing to the home.* Among the least common practices were teachers' reported efforts to extend writing to the home (see Table 4). For example, the two most common responses to assigning writing homework were never (23%) and several times a year (26%). Similar patterns were evident for asking parents to listen to something the child wrote at school (never = 29%; several times a year = 36%) or help their child write at home (never = 34%; several times a year = 28%). While all but a small percentage (7%) of teachers communicated with parents about their child's writing progress, slightly more than 2 out of every 3 teachers did this only several times a year.

*Connecting writing to reading and other content areas.* Although there was considerable variability in how often teachers reported using writing to support reading or students using writing in other content areas, these practices occurred at least weekly or more often with the majority of the teachers (see Table 4). It was less common for teachers to report using reading to support writing, with the largest proportion of teachers applying this practice either several times a year or several times a month.

There were grade-level differences in how frequently teachers reported using writing in other content areas,  $F(2, 157) = 7.08$ ,  $MSE = 18.56$ ,  $p = .000761$ . Grade 3 teachers were more likely to have their students write across the curriculum than Grade 2 teachers ( $p < .001$ ;  $ES = .69$ ; Grade 2:  $M = 3.0$ ,  $SD = 1.8$ ; Grade 3:  $M = 4.2$ ,  $SD = 1.4$ ).

## Discussion

By fourth grade, 2 out of every 3 children in the United States do not write well enough to meet classroom demands (Persky et al., 2003). This places these children at risk because writing is essential to educational and occupational success (National Commission on Writing, 2004, 2005). Concerns about students' writing development has led to calls to make writing a central element in the school reform movement, stressing that students need to write more, technology needs to be a more integral part of writing instruction, students' progress in writing needs to be monitored, and teachers need to be better prepared to teach writing (National Commission on Writing, 2003). While the need for better writing instruction cuts across all grade levels, we think that it is especially important to make improvements when children are first learning to write. It is difficult in later grades to overcome literacy problems that have their origin in the primary grades (Slavin et al., 1989). The development of policies and practices to improve writing instruction at any grade level, however, must be grounded in a clear understanding of how writing is currently taught. Without such information, it is hard to determine what needs to be done. The present study addresses this need, as it examined how writing is taught in Grades 1–3, drawing on a random sample of teachers from across the United States. We consider our findings in terms of their implications for future writing instruction for young children. This includes drawing new recommendations as well as examining the value of existing recommendations (specifically from the National Commission on Writing, 2003).

## *How Is Writing Taught in the Primary Grades?*

We hypothesized that primary grade teachers would report applying an eclectic approach to writing instruction, combining instructional procedures from both the process writing and skills approach. These are the most common approaches to teaching writing to young children in the United States (Applebee & Langer, 2006; Graham et al., 2002). With the process approach, considerable attention is placed on the act of writing, emphasizing extended opportunities to write for real audiences, creating a supportive writing environment, encouraging high levels of student interactions around writing, stressing personal responsibility for writing (including self-reflection and evaluation), and engaging in cycles of planning, translating, and revising (Pritchard & Hon-eycutt, 2006). The teaching of writing skills and strategies, however, is often personalized and provided when the need arises (although Nagin, 2003, for example, included sentence combining instruction in process writing). In contrast, writing skills instruction (sometimes referred to as *traditional instruction*) mainly emphasizes the explicit and systematic teaching of handwriting, spelling, sentence construction, and so forth, with writing itself receiving more or less attention depending upon the teacher.

Some professionals (see Freedman, 1993; Smith, 1994) have argued that approaches like process writing, which rely on informal and incidental learning methods, should not be combined with more traditional skills instruction. We anticipated, however, that teachers would favor a more balanced eclectic approach to writing instruction. This was the case when the teachers of young children reported how they taught the related skill of reading (see Baumann, Hoffman, Moon, & Duffy-Hester, 1998), described their beliefs about writing instruction (Graham et al., 2002), and indicated how they taught writing (Graham et al., 2003). The findings from the current study supported this prediction and corroborated the findings from the previous national survey of writing practices conducted by Graham et al. (2003).

In this study, almost 3 out of every 4 teachers indicated that they used a process approach combined with traditional skills instruction. These data taken alone are relatively weak, as process writing and traditional skill instruction were not defined in the survey. However, this combined approach was further reflected in the teachers' reported application of specific instructional practices. The typical teacher placed considerable emphasis on teaching basic writing skills, as spelling, grammar, capitalization, and punctuation skills were reportedly taught daily, with handwriting and sentence construction skills taught several times a week. The typical teacher also reported using a variety of practices common to the process writing approach. This included having students plan (at least weekly) and revise their compositions (at least several times a month), conference with and help other students with their writing (at least several times a month), share their writing with classmates (at least weekly), monitor their writing progress (at least weekly), choose their own writing topics (at least half the time), work at their own pace (at least half of the time), and use invented spellings (most of the time). Although Graham et al. (2003) did not ask teachers about three of these activities (i.e., sentence construction, student/student conferences, and teacher monitoring of students' progress), the teachers in the prior study indicated using six of these practices (i.e., grammar instruction, capitalization/punctuation instruction, student planning, student



sharing, student choice, and invented spelling) as frequently as teachers in this investigation. Teachers in the prior study, however, reported providing spelling instruction less often than teachers in this investigation (weekly vs. daily), but indicated that student revising (weekly vs. several times a month), peers helping each other (weekly vs. several times a month), and students working at their own pace (more than one half the time vs. one half the time) occurred more frequently. These differences are not large and may simply reflect differences in the samples surveyed.

The finding that most teachers applied both process and traditional skill instruction does not mean they applied each equally. Practices associated with traditional skill instruction occurred more often than those associated with the process writing approach (see above). We address the issue of how much emphasis teachers placed on process and skills as well as writing later (under *Additional Recommendations*).

It is interesting to note that teachers reported using most of the instructional practices that we identified. Ninety percent of the teachers reported using 3 out of every 4 practices at least sometime during the year (90% of the teachers used 19 of the 20 practices surveyed in Graham et al., 2003). There was, however, considerable variability in terms of how often a practice was applied (which was also the case in Graham et al., 2003). While we can only speculate on why some teachers readily apply a practice and others do not (e.g., teachers may differ in their views on the value of a particular technique), these findings are troublesome. It may not be enough to introduce teachers to new writing practices and encourage them to apply them. Efforts to reform writing instruction are likely to fall short, if little attention is devoted to how frequently practices are implemented. This needs to be the focus of both preservice as well as inservice professional development efforts.

We did not ask teachers participating in this study if they provided different instruction for different types of students, such as students with learning disabilities or students who were English language learners. This needs to be examined in future research.

### *Support for the Recommendations of the National Commission on Writing*

*Increase the amount of time students spend writing.* The findings from the present study address four recommendations, at least at the primary grade level, for writing instruction reform offered by the National Commission on Writing (2003). One of the commission's recommendations was to increase (i.e., double) the amount of time students spend writing. The median time teachers reported that their students spent writing each day was about 20 min (this involved writing material at least paragraph length or longer). In Graham et al. (2003), the median time was 30 min a day (but included any type of writing—not just paragraph length or longer). In both studies, a remarkably small percentage of the school day was devoted to writing in either study, providing support for the commission's recommendation. Two possible avenues for increasing amount of writing are to have children write more at home and write more often when working on other subjects (more than half of the participating teachers did this just once a week or less in this study).

The data from this study also provide support for a collateral recommendation regarding increased writing: Primary grade stu-

dents should spend more time writing expository text. The most common writing activities in the participating teachers' classes focused on narrative writing (stories, personal narratives, and poems), writing to communicate (letters), completing worksheets, and responding to material read. Expository writing activities, such as writing to inform or persuade, were much less common. However, recent intervention studies demonstrate that even struggling writers in second and third grade can learn to successfully carry out such writing assignments (Graham, Harris, & Mason, 2005; Harris, Graham, & Mason, 2006).

*Make technology a more integral part of teaching writing.* The National Commission on Writing (2003) also recommended that technology should be a more integral part of writing instruction. This includes enhancing the technology infrastructure for writing in schools (hardware and software), as well as training teachers and students to use existing and new technology. While we only examined one aspect of this recommendation—students' use of computers during the writing period—the findings are startling. Students did not use computers during the writing period in 42% of the participating teachers' classrooms, and they only used them several times a year in another 25% of cases. These findings are even more disquieting than the statistics from an earlier national survey (Graham et al., 2003) where 30% of primary grade teachers reported never using computers during the writing period, and 16% reported using them just several times a year. There is ample evidence that computers can enhance the quality of children's writing (Bangert-Drowns, 1993; Goldring, Russell, & Cook, 2003; Graham & Perin, 2007a, 2007c). Thus, the limited use of computer in this and the previous survey by Graham et al. (2003) reinforces the call to make technology a more integral part of writing instruction, at least for primary grade children. It must also be noted that the use of computers is surely related to the number of computers in each class or school. While efforts to increase both hardware and software in schools have been ongoing, they "are often inadequate and frequently unequal" (National Commission on Writing, 2003, p. 23).

*Improve the preparation of teachers.* Another recommendation of the National Commission on Writing (2003) was that teachers need to be better prepared to teach writing. On a positive note, most of the teachers in this study moderately agreed that they were effective teachers of writing, liked to teach writing, and liked to write. When asked about the quality of their college teacher certification program, 28% indicated that it did a very good to exceptional job in preparing them to teach writing. However, another 28% indicated that their preparation was poor or inadequate, with an additional 44% noting that it was adequate. Clearly there is room for improvement. Thus, these findings provide tentative support for a recommendation that teacher education programs need to do a better job of preparing certification students to teach writing to primary grade students.

*Monitor students' writing progress.* The National Commission on Writing (2003) further recommended that students progress in writing be monitored and that assessments are aligned with standards and the curriculum, involve actual writing (not multiple-choice items), and are fair as well as authentic. In this study, we only focused on assessment at the classroom level. Approximately, 2 out of every 3 teachers reported that they monitored their students' writing weekly and had students monitor how they were doing as well. Only 9% of the teachers reported that they



never monitored progress or only did it several times a year. While these findings do not address how progress was monitored or if such monitoring was aligned to standards and the curriculum, they do suggest that teachers view assessment as an important part of their writing program. Additional research is needed, however, to better understand what and how primary grade teachers' assess writing.

### *Additional Recommendations*

**Balance.** The findings from this study provide the basis for three additional recommendations for improving primary grade writing at the classroom level. One area of concern involves the balance between teaching basic writing skills (e.g., handwriting, spelling, grammar, punctuation, and capitalization), teaching writing strategies and processes (e.g., planning, text organization, and revising), and writing text. Students in the typical classroom in this study spent about 1 hr per day writing or receiving writing instruction (based on median statistics and teacher reports). Almost 50% of this time was spent on the teaching of basic writing skills, 35% on writing, and 16% on teaching planning and revising (the percentages for medians are virtually identical in Graham et al., 2003). As we noted earlier, students are not spending enough time writing connected text, and we would also argue that the teaching of planning and revising strategies and processes are not receiving enough emphasis. Explicitly teaching young students how to plan and revise has a powerful impact on improving their writing (see a recent meta-analysis by Graham, 2006a). Furthermore, the available data call into question the effectiveness of systematically teaching grammar and usage (Andrews et al., 2006; Graham & Perin, 2007a). However, close to one half of the time teachers reported spending on skills instruction involved the teaching of grammar and usage. Drawing on all of these data, we recommend that primary grade writing instruction needs to strike an appropriate balance between writing, teaching skills, and learning writing strategies and processes. While students must learn how to write letters fluently, correctly spell words, and so forth (see Graham & Harris, 2000, for an empirical and theoretical justification for this claim), writing text and learning the strategic processes involved in writing should not be shortchanged.

**Motivation.** The typical teachers in this study reported that they frequently used several practices that should have a positive impact on students' motivation for writing, including allowing students to work at their own pace (at least half of the time), sharing their writing with peers (at least weekly), and modeling their enjoyment or love for writing (at least weekly). Other motivational activities, such as publishing and teachers sharing their writing, occurred much less frequently. In contrast, investigations that have studied exceptional primary grade literacy teachers find that their classrooms are rich with activities and procedures designed to foster motivation for writing (see, e.g., Pressley et al., 2005, 2006). There is also some evidence to suggest that motivation for writing influences writing development (Graham, 2006b; Graham, Berninger, & Fan, 2007). Accordingly, we recommend that primary grade teachers place greater emphasis on fostering students love and enjoyment of writing.

**School/home connections.** As a group, the teachers in this study did not report making strong connections for writing between school and home. The majority of teachers reported assign-

ing writing homework once a month or less. In addition, most teachers rarely (several times a year or less) indicated they communicated with parents about their child's writing progress, asked parents to listen to something the child wrote, or asked the child to write at home with their parents. As a result, we recommend that primary grade teachers develop strong school/home relationship designed to foster students' writing development.

### *Summary*

The present study was based on the assumption that primary grade teachers are aware of the elements of their teaching and can relate this knowledge to questions about how they teach writing. Although there is evidence that teachers can provide an accurate description of their literacy practices (see, for instance, Bridge & Hiebert, 1985; DeFord, 1985; Pressley et al., 1996) and we have some preliminary evidence that teachers' responses to the items on this survey accurately reflect what they do in the classroom (see Lane et al., 2008; Olinghouse, 2008), the findings from the current study need to be replicated, as well as supplemented by research where practices are observed and not just reported.

It must also be acknowledged that teachers' responses may be colored by their susceptibility to respond in socially desirable ways (i.e., to indicate they are doing something they are not because the former response would reflect poorly on their teaching). While this did not appear to be the case for some practices (e.g., making school/home connections occur infrequently even though this is a commonly recommended practice in education), it does not rule out this possibility with other items. Thus, this supports the need for additional research that applies observational techniques to the study of the writing practices of primary grade teacher (the lack of such data is a limitation in the current investigation). Nevertheless, this study is the only available contemporary investigation that looks at primary grade writing instruction nationwide, providing needed information for determining what needs to be done to improve writing instruction for young children. Although a prior study by Graham et al. (2003) surveyed primary grade teachers throughout the United States, it focused much more narrowly on how teachers adapted instruction for struggling writers.

The fact that teachers were randomly selected nationwide, a 61% return rate was obtained, and there was evidence that the responders and nonresponders were similar on selective variables (grade taught, type of school, location of school, size of school, and expenditures for commercial materials per student) diminishes concerns about response bias. It is important to note, however, that these safeguards do not completely alleviate this concern, as other factors, such as their disposition toward the teaching of writing, may have distinguished responders from nonresponders.

It must further be noted that for the 41 eight-point Likert-type items, sampling error was  $\pm 4.9\%$  given the size of the population, the number of participants completing the survey, the expected variation in teacher responses, and a statistical confidence level of 95%. However, for items that involved a yes/no response (e.g., the application of specific writing assignment, such as story writing), the sampling error was larger ( $\pm 7.3\%$ ). In addition, considerable caution must be applied to any findings concerning specific grade-level effects, as the sampling error is even greater.

It is also important to note that the present study just focused on classroom practices. We did not examine other factors, such as



school-wide, district, or state policies, that shape writing instruction. Furthermore, we did not examine all possible aspects of classroom writing practices. This would be impossible in a survey study, as very few teachers would be willing to complete such a questionnaire. We attempted to address this issue by drawing on previous research as well as expert recommendations for teaching writing when developing items to assess classroom practices, but our findings and recommendations must be interpreted in light of these limitations.

In summary, we found that most primary grade teachers take an eclectic approach to writing instruction, combining elements from the two most common methods for teaching writing: process writing and skills instruction (although they generally place less emphasis on process writing). In addition, almost all teachers reported using most of the practices surveyed, but there was considerable variability between teachers in how often they applied each practice. This finding yields an important implication for reform efforts. Such efforts will need to go beyond teachers just learning and applying new procedures and focus on how frequently new practices are applied.

The findings from this study and to a lesser extent Graham et al. (2003) also provide some directions for what needs to be done to reform writing instruction. Primary grade students need to spend more time writing, including writing expository text. There needs to be a better balance between time spent writing, learning writing process and strategies, and teaching writing skills, with more emphasis placed on the first two elements. Teachers need to pay more attention to promoting students' love and motivation for writing. Stronger connections for writing between school and home need to be established. Computers also need to become a more integral part of the writing program at the primary levels. Finally, teacher education programs need to do a better job of professional development in the teaching of writing to young children.

## References

- Achieve, Inc. (2005). *Rising to the challenge: Are high school graduates prepared for college and work?* Washington, DC: Author.
- Andrews, R., Torgerson, C., Bevertson, S., Freeman, A., Locke, T., Low, G., et al. (2006). The effects of grammar teaching on writing development. *British Educational Research Journal*, 32, 39–55.
- Applebee, A., & Langer, J. (2006). *The state of writing instruction: What existing data tell us*. Albany, NY: Center on English Learning and Achievement.
- Bangert-Drowns, R. (1993). The word processor as an instructional tool: A meta-analysis of word processing in writing instruction. *Review of Educational Research*, 63, 69–93.
- Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based Writing-to-Learn interventions on academic achievement: A meta-analysis. *Review of Educational Research*, 74, 29–58.
- Baumann, J., Hoffman, J. V., Moon, J., & Duffy-Hester, A. M. (1998). Where are teachers' voices in the phonics/whole language debate? Results from a survey of U.S. elementary classroom teachers. *Reading Teacher*, 51, 636–650.
- Bridge, C. A., Compton-Hall, M., & Cantrell, S. C. (1997). Classroom writing practices revisited: The effects of statewide reform on writing instruction. *Elementary School Journal*, 98(2), 151–170.
- Bridge, C. A., & Hiebert, E. H. (1985). A comparison of classroom writing practices, teachers' perceptions of their writing instruction, and textbook recommendations on writing practices. *Elementary School Journal*, 86, 155–172.
- Christenson, S. L., Thurlow, M. L., Ysseldyke, J. E., & McVicar, R. (1989). Written language instruction for students with mild handicaps: Is there enough quantity to ensure quality? *Learning Disability Quarterly*, 12, 219–222.
- Culham, R. (2003). *6 + 1 traits of writing*. New York: Scholastic Professional Books.
- DeFord, D. (1985). Validating the construct of theoretical orientation in reading. *Reading Research Quarterly*, 20, 351–367.
- Diaper, D. (1989). *Knowledge elicitation: Principles, techniques, and application*. New York: Wiley.
- Dillman, D. (2000). *Mail and internet surveys*. New York: Wiley.
- Duke, N. K., & Bennett-Armistead, V. S. (2003). *Reading & writing information text in the primary grades*. New York: Scholastic Teaching Resources.
- Fisher, C. W., & Hiebert, E. H. (1990). Characteristics of tasks in two approaches to literacy instruction. *Elementary School Journal*, 91(1), 3–18.
- Freedman, A. (1993). Show and tell? The role of explicit teaching in the learning of new genres. *Research in the Teaching of English*, 27, 222–251.
- Gillet, J. W., & Beverly, L. (2001). *Directing the writing workshop*. New York: Guilford Press.
- Goldring, A., Russell, M., & Cook, A. (2003). The effects of computers on student writing: A meta-analysis of studies from 1992–2002. *Journal of Technology, Learning, and Assessment*, 2, 1–51.
- Graham, S. (2000). Should the natural learning approach replace traditional spelling instruction. *Journal of Educational Psychology*, 92, 235–247.
- Graham, S. (2006a). Strategy instruction and the teaching of writing. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (187–207). New York: Guilford Press.
- Graham, S. (2006b). Writing. In P. Alexander & P. Winne (Eds.), *Handbook of educational psychology* (pp. 457–477). Mahwah, NJ: Erlbaum.
- Graham, S., Berninger, V., & Fan, W. (2007). The structural relationship between writing attitude and writing achievement in young children. *Contemporary Educational Psychology*, 32, 516–536.
- Graham, S., & Harris, K. R. (2000). The role of self-regulation and transcription skills in writing and writing development. *Educational Psychologist*, 35, 3–12.
- Graham, S., Harris, K. R., Fink-Chorzempa, B., & MacArthur, C. (2002). Primary grade teachers' theoretical orientations concerning writing instruction: Construct validation and a nationwide survey. *Contemporary Educational Psychology*, 27, 147–166.
- Graham, S., Harris, K. R., Fink-Chorzempa, B., & MacArthur, C. (2003). Primary grade teachers' instructional adaptations for struggling writers: A national survey. *Journal of Educational Psychology*, 95, 279–292.
- Graham, S., Harris, K. R., & Mason, L. (2005). Improving the writing performance, knowledge, and motivation of struggling young writers: The effects of self-regulated strategy development. *Contemporary Educational Psychology*, 30, 207–241.
- Graham, S., & Perin, D. (2007a). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, 99, 445–476.
- Graham, S., & Perrin, D. (2007b). What we know, what we still need to know: Teaching adolescents to write. *Scientific Studies in Reading*, 11, 313–336.
- Graham, S., & Perin, D. (2007c). *Writing next: Effective strategies to improve writing of adolescent middle and high school*. Washington, DC: Alliance for Excellence in Education.
- Greenwald, E., Persky, H., Ambell, J., & Mazzeo, J. (1999). *National assessment of Educational progress: 1998 report card for the nation and the states*. Washington, DC: U.S. Department of Education.
- Harris, K. R., Graham, S., & Mason, L. (2006). Improving the writing, knowledge, and motivation of struggling young writers: Effects of

- self-regulated strategy development with and without peer support. *American Educational Research Journal*, 43, 295–340.
- Hayes, J. (1996). A new framework for understanding cognition and affect in writing. In M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1–27). Mahwah, NJ: Erlbaum.
- Lane, K., Graham, S., Harris, K. R., Little, L., Sandmel, K., & Brindle, M. (2008). *Story writing: The effects of self-regulated strategy development for second grade students with writing and behavioral difficulties*. Manuscript submitted for publication.
- Mariconda, B. (2001). *Teaching expository writing*. New York: Scholastic Professional Books.
- Nagin, C. (2003). *Because writing matters: Improving student writing in our schools*. San Francisco, CA: Jossey-Bass.
- National Commission on Writing. (2003, April). *The neglected "R": The need for a writing revolution*. Retrieved April 23, 2008, from [http://www.writingcommission.org/prod\\_downloads/writingcom/neglect-edr.pdf](http://www.writingcommission.org/prod_downloads/writingcom/neglect-edr.pdf)
- National Commission on Writing (2004, September). *Writing: A ticket to work . . . or a ticket out: A survey of business leaders*. Retrieved April 23, 2008, from [http://www.writingcommission.org/prod\\_downloads/writingcom/writing-ticket-to-work.pdf](http://www.writingcommission.org/prod_downloads/writingcom/writing-ticket-to-work.pdf)
- National Commission on Writing (2005, July). *Writing: A powerful message from state government*. Retrieved April 23, 2008, from [http://www.writingcommission.org/prod\\_downloads/writingcom/powerful-message-from-state.pdf](http://www.writingcommission.org/prod_downloads/writingcom/powerful-message-from-state.pdf)
- Nelson, N. K., Bahr, C. M., & Van Meter, A. M. (2004). *The writing lab approach to language instruction and intervention*. Baltimore: Brookes Publishing.
- No Child Left Behind Act of 2001, Pub. L. No. 107–110, 115 Stat. 1425 (2002).
- Olinghouse, N. G. (2008). *Modeling the writing development of second- and fourth-grade students*. Manuscript in preparation.
- Persky, H. R., Daane, M. C., & Jin, Y. (2003). *The nation's report card: Writing 2002*. (NCES 2003–529). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics.
- Pressley, M., Gaskins, I., Solic, K., & Collins, S. (2006). A portrait of Benchmark School: How a school produces high achievement in students who previously failed. *Journal of Educational Psychology*, 98, 282–306.
- Pressley, M., Mohan, L., Bogaert, L. R., & Fingeret, L. (2005). *How does Bennett Woods produce such high language arts achievement?* (Tech. Rep). East Lansing MI: Michigan State University, College of Education, Literacy Achievement Research Center.
- Pressley, M., Rankin, J., & Yokoi, L. (1996). A survey of instructional practices of primary grade teachers nominated as effective in promoting literacy. *Elementary School Journal*, 96, 363–384.
- Pressley, M., Raphael, L., Gallagher, J. D., & DiBella, J. (2004). Providence-St. Mel School: How a school that works for African-American students works. *Journal of Educational Psychology*, 96, 216–235.
- Pritchard, R., J., & Honeycutt, J. (2006). Process writing. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 275–290). New York: Guilford Press.
- Rankin-Erickson, J., & Pressley, M. (2000). A survey of instructional practices of special education teachers nominated as effective teachers of literacy. *Learning Disabilities Research & Practice*, 15, 206–225.
- Schaefer, L. M. (2001). *Teaching narrative writing*. New York: Scholastic Professional Books.
- Schultz, K., & Fecho, B. (2000). Society's child: Social context and writing development. *Educational Psychology*, 35, 51–62.
- Slavin, R., Madden, N., & Karweit, N. (1989). Effective programs for students at risk: Conclusions for practice and policy. In R. Slavin, N. Karweit, & N. Madden (Eds.), *Effective programs for students at risk* (pp. 21–54). Boston: Allyn & Bacon.
- Smith, C. (Moderator). (1994). *Whole language: The debate*. Bloomington, IN: ERIC Clearinghouse on Reading, English, and Communication.
- Spandel, V. (2005). *Creating writers through 6-trait writing assessment and instruction*. Boston: Allyn & Bacon.
- Stahl, S. A., Pagnucco, J. R., & Suttles, C. W. (1996). First graders' reading and writing instruction in traditional and process-oriented classes. *Journal of Educational Research*, 89(3), 131–144.
- Vaca, R., & Rasinski, T. (1992). *Case studies in whole language*. Fort Worth, TX: Harcourt Brace.
- Wharton-McDonald, R., Pressley, M., & Hampston, J. M. (1998). Literacy instruction in nine first-grade classrooms: Teacher characteristics and student achievement. *Elementary School Journal*, 99(2), 101–128.
- Wray, D., Medwell, J., Fox, R., & Poulson, L. (2000). The teaching practices of effective teachers of literacy. *Educational Review*, 52(1), 75–85.

Received October 4, 2007

Revision received May 6, 2008

Accepted May 6, 2008 ■



# Epistemological Beliefs' Contributions to Study Strategies of Asian Americans and European Americans

Marlene Schommer-Aikins  
Wichita State University

Marilyn Easter  
San Jose State University

Given the increasingly culturally diverse composition of students in American classrooms, understanding the nature of cultural differences is necessary to generate effective instructional interventions. This study examines the individual differences in epistemological beliefs, ways of knowing, study strategies, and academic performance among different cultural groups. These cultural groups include European Americans (Americans of European ancestry) and first and successive generations of Asian Americans. College junior and senior business majors completed instruments assessing epistemological beliefs, ways of knowing, and study strategies. Multivariate analyses of variances revealed significant differences among cultural groups in 5 study strategies (low anxiety, selecting main ideas, testing strategies, high motivation, and information processing), course grades, and reading comprehension. Regression analyses revealed that beliefs about learning speed, knowledge construction, characteristics of successful students, and separate knowing contributed to cultural differences. This study highlights the need to avoid strong stereotyping and to consider individual differences in the classroom.

**Keywords:** epistemological beliefs, cultural differences, study strategies, reading comprehension, Asian American

Since 1990, students' epistemological beliefs (beliefs about the nature of knowledge and learning) have been studied using a multidimensional paradigm (Schommer, 1990). Within this paradigm, researchers have examined beliefs about the structure of knowledge, the sources of knowledge, the certainty of knowledge, the speed of learning, and the ability to learn (Buehl, Alexander, & Murphy, 2002; Hofer, 2000; Jehng, Johnson, & Anderson, 1993). The accumulating evidence indicates that epistemological beliefs predict comprehension, metacomprehension, and academic performance (Schommer-Aikins, 2004). In this study, we compare the epistemological beliefs and study strategies of first and successive generations of Asian Americans and European Americans (Americans of European ancestry). We examine how these individual difference variables may contribute to cultural differences.

It is important to understand cultural differences through the lens of individual differences. To state that students in the class perform differently because they are from different cultures is insufficient. Instead, we should ask, "Specifically what about the cultures is making the difference?" Cultures differ in beliefs, norms, and behavior. Our focus is the difference in beliefs among cultures. We argue that these differences in beliefs (epistemology) are likely to lead to differences in cognition (study strategies) and subsequently to differences in academic performance.

We further argue that instructors may misinterpret students' behavior if they focus on cultural stereotypes rather than individual differences. For example, if students do not ask questions or participate in classroom discussions, this silence is open to the instructor's interpretation. Silence could be interpreted as a sign of understanding (based on a positive stereotype), or silence could be interpreted as a sign of lack of understanding or indifference (based on a negative stereotype; Chang & Demyan, 2007; Ng, Lee, & Pak, 2007). If instructors rely on stereotypes rather than on monitoring student understanding in multiple ways, students may not receive the academic support that they need.

## Consequences of Stereotypes

This study is an initial attempt to test our assertions. We tested for differences in beliefs and study strategies between cultures. We then tested the notion that epistemological beliefs and study strategies contribute to differences in academic performance.

Some cultural stereotypes can be misleading—for example, the "model minority." The U.S. news media assigned the term model minority to successful Asian Americans, presumably as a compliment and a source of inspiration to others (Kao, 1995; Tasaki, 2001), but the residual consequences to Asian Americans have been mixed (Lee, 1994). On the surface, this stereotype might seem advantageous. When Asian Americans are seen as hardworking and valuing education, others assume that they will excel in most academic subjects. However, this stereotype is an overgeneralization that is difficult to cope with if one does not meet the projected expectations.

For example, Lee (1994) portrays the burden and inappropriateness of the model-majority stereotype among Asian American high school students. First, it is presumptuous to categorize all Asian Americans into a single group. Asian Americans come from dif-

---

Marlene Schommer-Aikins, Department of Counseling, Educational and School Psychology, Wichita State University; Marilyn Easter, Department of Marketing, San Jose State University.

We thank Doris Burgert for her careful comments and editing of several drafts of this article.

Correspondence concerning this article should be addressed to Marlene Schommer-Aikins, College of Education, Hubbard Hall Room 320, Wichita State University, Wichita, KS 67260-0123. E-mail: marlene.schommer-aikins@wichita.edu

ferent countries, different socioeconomic groups, and different parenting styles. Second, Asian Americans are aware of the stereotype and may react differently to this characterization. In Lee's ethnographic study, four different student reactions were identified. Students reacted by doing one or more of the following: (a) distancing themselves from other Asians, (b) choosing less desirable careers, (c) interpreting a less-than-perfect academic performance as a failure, or (d) doing just enough academic work to get by.

In Lee's (1994) study, some students identified themselves as Korean. They distanced themselves from other Asians and were encouraged by their parents to emulate upper-middle-class European Americans outside the home to achieve success. Yet, at the same time, their parents encouraged them to maintain their Korean identity within the home and within Korean community events. Despite the positive attitude toward success, Korean students in this sample had varying levels of academic success. In response to this varying academic success, the higher achievers willingly helped the lower achievers.

The academic behavior of students who identified themselves as Asian was most consistent with the stereotype of the quiet, hard-working student (Lee, 1994). Their parents taught them that doing well in school was the key to success. Despite that optimistic view of school, Asian-identified students anticipated discrimination in the United States, particularly because of their accents. Students would often select careers that were less desirable to avoid rejection on the basis of language. For example, students who truly wanted to study law might go into engineering instead to avoid issues about accents. In Lee's (1994) study, although these students valued hard work and education, their academic achievement ranged from high to low.

Students who identified themselves as Asian American considered themselves neither exclusively Asian nor American (Lee, 1994). Rather, they had a new identity that was a blend of Asian and American. They also believed in hard work and doing well in school to fight racism. They were outspoken about racism by speaking out against the model-minority stereotype. For example, they argued that if Asian American students were not at the very top of their math class, they were considered failures. This put undue pressure on students who under other circumstances would have been considered to be performing well.

Asian American students who identified themselves as New Wavers took this term from New Wave music (Lee, 1994). These high school students made their identity and group membership obvious with their spiked hair and black clothes. Unlike other Asian Americans, they did not define success through school performance. Indeed, they blatantly disrespected academic rules to assume an identity that was the opposite of the expected stereotype, doing just enough work to pass their courses. Notions of future careers were vague. Allegiances were aligned more to their peers than to the aspirations offered by their parents or the school system. Aspirations were aimed more at physical prowess or beauty than academic achievement. Indeed, academic achievers were seen as "nerds," a label they did not want to endorse by actually doing well in school. In summary, Lee's work suggests that Asian American students' identities, motivation, and academic achievement varies greatly.

Finally, there is evidence that academic performance among cultural groups is not consistent with the model-minority hypoth-

esis when the academic tasks require creativity and open-ended thinking. Recently, Wang and Lin (2005) highlighted that cross-cultural studies on American and Chinese students regarding mathematics learning may be misleading. For example, Chinese students routinely outperform U.S. students in computational skills and routine problem-solving skills in mathematics. Less well known is that U.S. students typically outperform Chinese students on open-ended and creative problem-solving tasks (Cai, 1997, 1998, 2000). Wang and Lin attempted to generate potential explanations for the mixed findings; the dominant explanation was that classifying participants as Asian American masks a large variability within that group. There is immense diversity in the nation of origin, as well as language and social adjustment. Furthermore, there is a disparity between Asian Americans' mathematical performance and verbal performance. Wang and Lin suggested examining more individual difference variables within and between cultures and exploring more specific performance measures.

### Epistemically Related Beliefs

In this study we focus on an array of beliefs: some that are traditionally considered epistemological, such as belief in the certainty and structure of knowledge, and some that have only recently been considered intimately related to epistemology, such as belief in ways of knowing (Schommer-Aikins, 2004). Therefore, we refer to the beliefs examined in this study as *epistemically related beliefs*, which include beliefs about the nature of knowledge and learning (epistemological beliefs, Schommer, 1990) and ways of knowing (Belenky, Clinchy, Goldberger, & Tarule 1986). Epistemically related beliefs are thought to guide students' choices of study strategies and subsequently their academic performance. One's culture may contribute to epistemically related beliefs (Schommer-Aikins, 2004). For example, the more children are given an opportunity to question and discuss important issues with their parents, the more likely they are to believe that learning is a gradual process. Moreover, the more postsecondary education adults receive, the more likely they are to believe knowledge is tentative and complex (Schommer, 1994). Hence, both individual and larger, cultural perspectives likely inform epistemically related beliefs.

Perry's (1968) work on students' views about education initiated a general interest in epistemological beliefs. His interviews with Harvard undergraduate men led him to theorize that freshmen enter college believing that knowledge is simple, certain, and handed down by authority. In contrast, Perry found that seniors believed that knowledge is complex, tentative, and derived from reason based on evidence. Although many researchers developed lines of epistemic research closely aligned to Perry's thinking (e.g., Baxter Magolda, 1987; Kitchener & King, 1989; Ryan, 1984), two lines of research forged other paths of study.

One epistemic path was forged by Belenky et al. (1986), who were concerned that Perry's work was based mainly on a male population. Hence, in contrast to Perry, they conducted extensive interviews with women, from whom they theorized epistemic positions known as *ways of knowing*. More recent articulation of their work (Clinchy, 2002) focuses on two major epistemic positions, *separate knowing* and *connected knowing*.

Research indicates that both ways of knowing contribute to deeper thinking, each taking a different path to meet similar goals.



Individuals who emphasize separate knowing take the “devil’s advocate” approach. They doubt and question first, then attempt to embrace new ideas. Individuals who emphasize connected knowing take an empathic approach. They try to take on others’ roles to see the world from their perspective and to understand first, then engage in more evaluative thinking.

Studies (Clinchy, 2002; Galotti, Blythe, Ainsworth, Lavin, & Mansfield, 1999) have found these ways of knowing are gender related in the United States; men have a propensity toward separate knowing, and women have a propensity toward connected knowing. In other words, men tend to play the devil’s advocate, whereas women are more likely to think empathically. This does not imply, however, that these ways of knowing are gender specific. Indeed, individuals are arguably capable of both ways of knowing. Clinchy (2002) hypothesized that the most advanced thinkers have learned to use each way of knowing adaptively to meet situational demands.

Another epistemic path that moved beyond Perry’s work was forged by Schommer (1990). On the basis of a synthesis of research from 1968 to 1988 (e.g., Baxter Magolda, 1987; Dweck & Bempechat, 1983; Kitchener & King, 1981; Perry, 1968; Ryan, 1984; Schoenfeld, 1983), Schommer proposed the idea of an epistemological belief system. In this case, *system* means that personal epistemology is composed of a number of more or less independent belief dimensions; *independent* means that these beliefs do not necessarily mature at the same rate. These beliefs included the stability (from unchanging to evolving), structure (from isolated facts to integrated complexities), and source of knowledge (from handed down by authority to derived from reason and evidence); the speed of learning (from quick or not at all to gradual); and the ability to learn (from fixed at birth to incremental). The implication is that to understand individuals more completely, it is important to assess all of their beliefs. As a result of these recent developments in measurement, researchers continue to find important links between epistemological beliefs and various aspects of learning. Epistemological beliefs predict comprehension, metacomprehension, and interpretation of information (Schommer-Aikins, 2004).

The consequences of epistemological beliefs are arguably subtle because many of their effects are mediated by other variables (Bendixen & Rule, 2004; Schommer-Aikins, 2004). For example, if students believe that knowledge is structured as isolated facts, they will likely study by memorizing lists of concepts. In the classroom, students will recall words and/or definitions, but they will lack applicable or contextual understanding. To date there are few studies on the mediating effects of epistemological beliefs (e.g., Schommer, Crouse, & Rhodes, 1992; Schommer-Aikins, Duell, & Hutter, 2005; Schreiber & Shinn, 2003). Furthermore, these studies have focused on fairly homogeneous populations.

### Cross-Cultural Study of Epistemically Related Beliefs

Limited research has examined the multidimensional epistemological belief system in cross-cultural studies. Some researchers have explored the specific dimensions that constitute the personal epistemology of Chinese students and have found epistemological belief factors that merge (e.g., simple–certain; Qian & Alvermann, 2000) or factors composed of different nuances (e.g., omniscient authority shifting on factors; Chan & Elliott, 2002). The precise

nature of epistemological beliefs within cultures, as well as between them, remains unknown.

Qian and Pan (2002) compared American and Chinese 11th and 12th graders. They found that Chinese students were more likely to believe in simple knowledge and innate ability than American students. Karabenick and Moosa (2005) similarly found that Middle Eastern (Omani) students were more likely to believe that knowledge is simple and certain than American students. Omani male students were more accepting of omniscient authority than Omani female students were. No gender differences regarding epistemic beliefs were found with the American students in Karabenick and Moosa’s study.

In our work, comparisons of epistemological beliefs were made between European Americans (EUROA), first-generation Asian Americans (FIRSTAA), and beyond-first-generation Asian Americans (BEYONDAA). Distinctions between first-generation and beyond-first-generation Asian Americans were made because cultural differences from the mainstream are typically reduced in subsequent generations (Rozin, 2003). We investigated the roles that epistemological beliefs and study strategies play in academic performance involving open-ended and creative responses, because investigations of these types of academic tasks are limited in cross-cultural studies. We hypothesized that epistemological beliefs, ways of knowing, and study strategies are three arrays of individual difference variables that account for cultural differences.

In this study four hypotheses were tested:

*Hypothesis 1:* Cultural groups will differ in their epistemological beliefs and ways of knowing. Specifically, EUROA will have higher scores on epistemological beliefs.

*Hypothesis 2:* Study strategy differences will be explained by epistemological beliefs, ways of knowing, and/or gender. Specifically, if study strategy differences are found among cultural groups, epistemological beliefs and ways of knowing will contribute to those differences.

*Hypothesis 3:* Cultural groups will differ in nonmathematical academic performance. Specifically, EUROA will score higher on academic performances that require open-ended and creative responses.

*Hypothesis 4:* Academic performance differences between groups will be partially explained by study strategies, epistemological beliefs, ways of knowing, and gender. Specifically, at least part of the variance due to cultural differences will be accounted for by individual differences.

### Method

#### Participants

Participants were 264 college juniors ( $n = 151$ ) and seniors ( $n = 113$ ) enrolled in a business communication class at a West Coast U.S. university. There were 88 FIRSTAA (57% female, 43% male), 90 BEYONDAA (57% female, 43% male), and 78 EUROA (51% female, 49% male). The average age was 24 years old ( $SD = 5.61$ ), with a range of 19 to 54 years old. The youngest group was BEYONDAA ( $M = 22.54$ ,  $SD = 3.21$ ). The next



youngest group was FIRSTAA ( $M = 24.31$ ,  $SD = 4.47$ ). The oldest group was EUROA ( $M = 25.97$ ,  $SD = 7.99$ ). There was one significant difference in age between BEYONDAA and EUROA:  $F(2, 247) = 7.89$ ,  $p < .01$ ,  $\eta^2 = .06$ . Students were asked if English was their first language; 26.1% of FIRSTAA, 51.1% of BEYONDAA, and 94.9% of EUROA responded "yes." Students received extra credit for their participation in the study.

### Instruments

**Study strategies.** Study strategies were assessed using the Learning and Study Strategies Inventory (LASSI; Weinstein & Palmer, 2002). This instrument assesses study strategies related to technique, motivation, and self-regulation. Subtests and sample items include the following: low anxiety about academic performance (Low Anxiety), "Worrying about doing poorly interferes with my concentration on tests" (reverse scored); positive attitude toward school (Positive Attitude), "I feel confused and undecided as to what my educational goals should be" (reverse scored); concentration (Concentration), "I concentrate fully when studying"; information processing (Information Processing), "I translate what I am studying into my own words"; high academic motivation (High Motivation), "I set high standards for myself in school"; selecting main ideas (Selecting Main Ideas), "I have difficulty identifying the important points in my reading"; self-testing (Self-Testing), "I stop periodically while reading and mentally go over or review what was said"; use of study aids (Study Aids), "I use special helps, such as italics and headings, that are in my text"; test strategies (Test Strategies), "I have difficulty adapting my studying to different types of courses" (reverse scored); and time management (Time Management), "When I decide to study, I set aside a specific length of time and I stick to it." Subtests were composed of eight items each with the exception of Selecting Main Ideas, which was composed of five items. This is a field-tested instrument with over 1,000 students represented. Weinstein and Palmer (2002) reported that Cronbach's alpha for subscales ranged from .73 to .89.

**Epistemically related beliefs.** Ways of knowing were measured with the Attitude Toward Thinking and Learning Survey (ATTLS; Gallotti et al., 1999). This 20-item questionnaire generates unique scores for Separate Knowing and Connected Knowing composed of 10 items each. Students respond to statements such as "I try to think with people instead of against them," and "I like playing the devil's advocate arguing the opposite of what someone is saying." They were measured on a Likert scale from 1 (*strongly disagree*) to 7 (*strongly agree*). Gallotti et al. (1999) reported Cronbach's alpha from each scale of .83 (Connected Knowing) and .77 (Separate Knowing). Details of psychometric properties and instrument development can be seen in their report.

Epistemological beliefs were measured using the Kardash Epistemological Belief Scale (Kardash & Wood, 2000; Wood & Kardash, 2002). This questionnaire evaluates student beliefs about the speed of knowledge acquisition (Speed), composed of eight items ranging from "quick all-or-none learning" to "learning is a gradual process that takes effort"; the structure of knowledge (Structure), composed of 11 items ranging from "knowledge is organized as isolated facts" to "knowledge is structured in integrated, complex, and sometime ambiguous concepts"; knowledge construction and modification (Construction), composed of 11 items ranging from

"knowledge is certain and passively received" to "knowledge is constantly evolving and it takes active modification and personal construction to learn"; characteristics of successful students (Success), composed of five items ranging from "good learners have innate ability and are able to memorize and learn with ease" to "although good learners have acquired study skills, learning still takes time and effort"; and attainability of truth (Truth), composed of three items ranging from "there is objective truth attainable by scientists" to "there are seldom single answers and facts/truths should be open to question." Students responded to statements such as "Usually, if you are ever going to understand something, it will make sense to you the first time" (Speed, reverse scored); "I like information to be presented in a straightforward fashion. I don't like having to read between the lines" (Structure, reversed score); and "Wisdom is not knowing the answers but knowing how to find the answers" (Construction). Items were measured on a Likert scale from 1 (*strongly disagree*) to 5 (*strongly agree*). Kardash and Wood (2000) reported that Cronbach's alpha for these belief scales ranged from .54 to .74. Details of psychometric properties and instrument development can be seen in Wood and Kardash (2002).

In a previous study (Easter & Schommer-Aikins, 2004), all of the described instruments were pilot-tested for this population. Students were instructed either to ask the researcher questions or to write comments next to items that were unclear. Students indicated that all instruments were easy to understand.

These beliefs and study strategies, as conceived in existing studies, have been shown to predict academic performance in U.S. schools. Because this study was conducted in U.S. schools, scores were calculated using formulas from the authors of the instruments. This does not suggest that we believe that these scores reflect cultural nuances of each group. Rather, these beliefs likely are important for success in the U.S. schools.

**Academic performance.** Two measures of academic performance were obtained. All business communication students were required to take the Reading Comprehension test from the NCS Pearson (2000) Reading and Arithmetic Indexes-12. This test assesses word decoding, phrase comprehension, sentence comprehension, and paragraph comprehension. For example, a typical item for sentence comprehension would be as follows: "A man in Texas, trying to call Iowa City, dialed the wrong number." Students then select from several options such as the following: (a) "The man dialed the wrong number in Iowa"; (b) "The man, while trying to dial Iowa City, dialed the wrong number"; (c) "The man called Texas instead of Iowa City"; and (d) "The man could not call Texas or Iowa City." This instrument assesses basic reading commonly required for entry-level employment, including office/clerical, vocational, and manufacturing jobs. Scores can range from 1 to 72. Test-retest reliabilities for this reading test range from .86 to .93 (NCS Pearson, 2000). Details of psychometric properties are reported in the NCS Pearson (2000) manual for the test.

Student course grades for the business communication class were also used as a measure of academic performance. Throughout the semester, instructors provided students with feedback on more than 30 pages of written assignments. Instructors also evaluated students through midterm and final exams, weekly quizzes, in-class writing exercises, and verbal discussions. All instructors used



consistent grading criteria for assignments. Final grades for this study were coded numerically from 0 (F) to 12 (A).

**Demographics.** Basic demographic information was assessed as well. Students were asked to report their age, gender, and ethnicity. They were also asked if English was their first language—that is, the primary language they used regularly.

### Procedure

Surveys were administered by classroom instructors over a 4-week period. The sequence of the survey's administration was counterbalanced to control for possible order effects. Surveys were completed at the beginning of each class, before any other activities. Half of the class received one survey (e.g., ATTLS), and the remaining students in the class completed a different survey (e.g., LASSI). Each survey took approximately 15–30 min to complete.

### Results

**Scores for each measure.** Scores were calculated for each survey instrument, and interitem correlations of Cronbach's alpha were calculated for each cultural group. Cronbach's alpha for this study was similar to previous studies (see Wood & Kardash, 2002). In the study being reported, Cronbach's alpha for LASSI ranged from .59 to .81, ways of knowing ranged from .60 to .81, and epistemological beliefs ranged from .50 to .82. Cronbach's alpha for each measure broken down by cultural group are shown in Table 1. For all scores in this study, higher scores represent better support for learning. For example, a high score on the Low Anxiety scale means that the student can cope with anxiety well. A high score on Speed means that the student believes learning may require more time to complete.

**Epistemological belief differences.** To test the hypothesis that cultural groups differ in their epistemological beliefs, we conducted a two-way multivariate analysis of variance (MANOVA; Cultural Group  $\times$  Gender) using epistemological beliefs scores as the dependent variables. Wilks's lambda multivariate statistic was

Table 1  
Cronbach's  $\alpha$  for Epistemological Beliefs and Ways of Knowing

Variable	FIRSTAA	BEYONDAA	EUROA
Structure	.50	.53	.64
Construction	.53	.51	.50
Success	.54	.67	.67
Truth	.75	.52	.82
Speed	.64	.74	.58
Connected Knowing	.80	.81	.79
Separate Knowing	.71	.69	.74
Low Anxiety	.87	.82	.87
Positive Attitude	.67	.71	.84
Concentration	.76	.78	.81
Information Processing	.81	.78	.82
High Motivation	.79	.81	.87
Self Testing	.80	.81	.79
Select Main Ideas	.86	.85	.91
Study Aids	.72	.59	.63
Time Management	.82	.81	.85
Test Strategies	.79	.79	.83

Note. FIRSTAA = first-generation Asian Americans; BEYONDAA = beyond-first-generation Asian Americans; EUROA = Euro-Americans.

significant for cultural groups:  $F(10, 466) = 6.59, p < .001, \eta^2 = .12$ . Univariate analyses indicated two significant main effects for cultural groups in Speed,  $F(2, 236) = 18.39, p < .001, \eta^2 = .14$ , and Structure,  $F(2, 236) = 11.80, p < .001, \eta^2 = .09$ . Follow-up Tukey post hoc tests indicated the EUROA had significantly higher scores for Speed compared to both FIRSTAA,  $F(1, 158) = 40.99, p < .001, \eta^2 = .20$ , and BEYONDAA,  $F(1, 159) = 19.20, p < .001, \eta^2 = .11$ . EUROA also had significantly higher for Structure compared to both FIRSTAA,  $F(1, 157) = 16.09, p < .001, \eta^2 = .09$ , and BEYONDAA,  $F(1, 158) = 18.39, p < .001, \eta^2 = .10$ . In other words, the EUROA group had a stronger belief that learning takes time and that knowledge is organized as a complex network. There were no significant differences between the two Asian American groups (FIRSTAA and BEYONDAA). Descriptive statistics for epistemological beliefs are shown in Table 2.

**Ways of knowing differences.** To address the hypothesis that cultural groups differ in their ways of knowing, another two-way MANOVA (Cultural Group  $\times$  Gender) was conducted with Separate Knowing and Connected Knowing scores as dependent variables. Wilks's lambda multivariate statistic was significant for gender,  $F(2, 249) = 16.32, p < .001, \eta^2 = .12$ . Women had a stronger belief in connected knowing (women,  $M = 55.13, SD = 8.10$ ; men,  $M = 51.37, SD = 8.28$ ),  $F(1, 250) = 14.22, p < .001, \eta^2 = .05$ ; and men had a stronger belief in separate knowing (women,  $M = 43.85, SD = 7.78$ ; men,  $M = 46.21, SD = 8.28$ ),  $F(1, 250) = 4.99, p < .03, \eta^2 = .02$ .

**Study strategy differences.** To address the hypothesis that cultural groups differ in their study strategies, a two-way MANOVA (Cultural Group  $\times$  Gender) was conducted with the LASSI scores as dependent variables. Wilks's lambda multivariate statistic was significant for both cultural groups,  $F(20, 450) = 2.50, p < .001, \eta^2 = .10$ , and gender,  $F(10, 225) = 3.80, p < .001, \eta^2 = .15$ . Univariate analyses indicated that cultural groups differed on Low Anxiety,  $F(2, 234) = 4.54, p < .01, \eta^2 = .04$ ; Information Processing,  $F(2, 234) = 4.08, p < .05, \eta^2 = .03$ ; High Motivation,  $F(2, 234) = 6.44, p < .01, \eta^2 = .05$ ; Selecting Main Ideas,  $F(2, 234) = 5.18, p < .01, \eta^2 = .04$ ; and Testing Strategies,  $F(2, 234) = 4.71, p < .01, \eta^2 = .04$ . Follow-up Tukey post hoc tests indicated that EUROA, compared to FIRSTAA, were better able to control their anxiety about school,  $F(1, 156) = 11.82, p < .001, \eta^2 = .07$ ; select main ideas,  $F(1, 156) = 7.56, p < .001, \eta^2 = .05$ ; and prepare for tests,  $F(1, 156) = 8.07, p < .001, \eta^2 = .05$ . Compared to BEYONDAA students, EUROA students were better able to select main ideas,  $F(1, 165) = 10.87, p < .001, \eta^2 = .06$ ; use information processing strategies,  $F(1, 165) = 10.73, p < .001, \eta^2 = .06$ ; and be academically motivated,  $F(1, 165) = 18.44, p < .001, \eta^2 = .10$ . There were no statistically significant differences between the two Asian American groups.

Univariate analyses for gender comparisons indicated that women, overall, were better able to use study aids (women,  $M = 26.03, SD = 4.49$ ; men,  $M = 23.68, SD = 5.07$ ),  $F(1, 234) = 13.70, p < .01, \eta^2 = .06$ ; better able to manage their time (women,  $M = 26.00, SD = 6.04$ ; men,  $M = 24.15, SD = 5.77$ ),  $F(1, 234) = 8.79, p < .01, \eta^2 = .04$ ; and had a better attitude toward school (women,  $M = 32.05, SD = 4.83$ ; men,  $M = 30.75, SD = 4.46$ ),  $F(1, 234) = 5.62, p < .05, \eta^2 = .02$ . Men were better able to control their anxiety about school (women,  $M = 25.02, SD = 7.60$ ;

Table 2  
*Epistemological Beliefs by Cultural Groups*

Belief	Cultural group		
	EUROA <i>M (SD)</i>	FIRSTAA <i>M (SD)</i>	BEYONDAA <i>M (SD)</i>
Speed <sup>a</sup>	4.18 (0.35)	3.72 (0.52)	3.86 (0.54)
Structure <sup>a</sup>	2.85 (0.47)	2.56 (0.41)	2.56 (0.38)
Construction	3.68 (0.36)	3.65 (0.36)	3.67 (0.32)
Success	3.27 (0.75)	3.18 (0.69)	3.33 (0.75)
Truth	3.30 (1.03)	3.06 (0.90)	3.23 (0.81)

Note. EUROA = Euro-Americans; FIRSTAA = first-generation Asian Americans; BEYONDAA = beyond-first-generation Asian Americans.

<sup>a</sup>EUROA had significantly higher belief scores compared to both Asian American groups.

men,  $M = 27.47$ ,  $SD = 6.01$ ),  $F(1, 234) = 6.25$ ,  $p < .05$ ,  $\eta^2 = .03$ . Descriptive statistics for the LASSI are shown in Table 3.

*Epistemological beliefs and ways of knowing as predictors of study strategies.* To test the hypothesis that epistemological beliefs and ways of knowing explain some of the cultural differences in study strategies, regression analyses were conducted. Each LASSI variable that was significant (for cultural differences) in the previous findings served as a dependent variable. In the first block of variables to enter the equation, epistemological beliefs, ways of knowing, and gender competed for entry, if significant at the .05 level. The second block of variables that entered the equation, if significant at the .05 level, was the appropriate contrast, either EUROA versus FIRSTAA or EUROA versus BEYONDAA. A summary of these regressions is shown in Table 4.

For the EUROA versus FIRSTAA comparison, differences between cultural groups on LASSI variables were accounted for by epistemological beliefs, ways of knowing, and gender. Specifically, Speed and gender predicted Low Anxiety. The more students believed learning is gradual, the less anxiety they reported. Men reported less anxiety than women did.

Speed, Separate Knowing, and Success predicted Selecting Main Ideas. The more students believed that learning is gradual, believed in separate knowing, and believed that success is related to hard work, the more likely they were to report that they could identify main ideas.

Speed and Separate Knowing predicted Testing Strategies. The more students believed that learning is gradual and believed in separate knowing, the more likely they were to report preparing for tests strategically. Cultural differences were no longer significant.

For the EUROA to BEYONDAA comparison, differences in LASSI variables were partially accounted for by epistemological beliefs and ways of knowing. A summary of these regressions is shown in Table 5. Specifically, Speed predicted Selecting Main Ideas. The more students believed that learning is gradual, the more likely they were to report being able to select main ideas.

Speed and the culture contrast predicted High Motivation. The more students believed learning is gradual, the more likely they were to report being academically motivated. EUROA students reported higher academic motivation than BEYONDAA students.

Construction, Separate Knowing, and culture contrast predicted Information Processing. The more students believed that certain knowledge is actively acquired and believed in separate knowing,

the more likely they were to report using information processing study strategies. EUROA students reported higher use of information processing strategies than BEYONDAA students did.

*Academic performance differences.* To test the hypothesis that there are differences between cultural groups in academic performance, two-way ANOVAs were conducted with the final course grade as a dependent variable in one analysis and reading scores as the dependent variable in another analysis. Gender and cultural groups served as independent variables. With the final course grade as the dependent variable, there was a main effect for cultural groups,  $F(2, 225) = 6.51$ ,  $p < .01$ ,  $\eta^2 = .05$ . The follow-up Tukey post hoc test indicated that EUROA students outperformed both FIRSTAA students,  $F(1, 147) = 10.37$ ,  $p < .01$ ,  $\eta^2 = .07$ , and BEYONDAA students,  $F(1, 151) = 9.61$ ,  $p < .01$ ,  $\eta^2 = .06$ . There was also a main effect for gender, in which women outperformed men (women,  $M = 8.73$ ,  $SD = 2.72$ ; men,  $M = 8.01$ ,  $SD = 3.29$ ),  $F(1, 225) = 4.51$ ,  $p < .05$ ,  $\eta^2 = .05$ .

With reading scores as the dependent variable, there was a main effect for the cultural groups,  $F(2, 233) = 5.44$ ,  $p < .01$ ,  $\eta^2 = .05$ . The follow-up Tukey post hoc tests indicated that EUROA students outperformed FIRSTAA students,  $F(1, 155) = 11.91$ ,  $p < .001$ ,  $\eta^2 = .07$ . There were no statistically significant differences between the two Asian American groups. Descriptive statistics for academic performance are shown in Table 6.

*Individual differences as predictors of academic performance.* To test the hypothesis that epistemological beliefs, ways of knowing, and study strategies account for cultural group differences in academic performance, hierarchical multiple regression analyses were conducted with the final grade as the criterion variable in one analysis and reading score as the criterion variable in another analysis. In the first block of variables to enter the equation, epistemological beliefs, ways of knowing, LASSI, English as first language, and gender were allowed to enter the equation if they were significant at the .05 level. The second block of variables to enter the equation was the appropriate contrast, either EUROA versus both Asian American groups or EUROA versus FIRSTAA, if significant at the .05 level. A summary of these regressions is shown in Table 7.

Table 3  
*LASSI Scores by Cultural Groups*

Study strategy	Cultural group		
	EUROA <i>M (SD)</i>	FIRSTAA <i>M (SD)</i>	BEYONDAA <i>M (SD)</i>
Low Anxiety <sup>a</sup>	28.11 (7.23)	24.59 (7.32)	26.24 (6.27)
Positive Attitude	32.07 (5.44)	31.27 (4.42)	30.97 (4.29)
Concentration	28.28 (5.19)	26.30 (5.10)	27.16 (4.98)
Information Processing <sup>b</sup>	28.55 (5.54)	28.30 (5.30)	26.43 (4.94)
High Motivation <sup>b</sup>	32.73 (5.58)	30.98 (4.82)	29.88 (4.94)
Self Testing	23.68 (5.55)	24.45 (5.77)	23.52 (5.50)
Selecting Main Ideas <sup>a,b</sup>	30.82 (6.04)	28.22 (5.53)	28.36 (5.09)
Study Aids	24.30 (5.00)	25.79 (5.56)	24.57 (4.76)
Time Management	25.22 (6.84)	25.04 (5.92)	25.47 (5.40)
Testing Strategies <sup>a</sup>	30.44 (5.27)	27.81 (5.45)	29.17 (4.83)

Note. LASSI = Learning and Study Strategies Inventory; EUROA = Euro-Americans; FIRSTAA = first-generation Asian Americans; BEYONDAA = beyond-first-generation Asian Americans.

<sup>a</sup>EUROA scored significantly higher scores than FIRSTAA.

<sup>b</sup>EUROA scored significantly higher scores than BEYONDAA.



Table 4

*Summary of Regression Analyses: Epistemological Beliefs as Predictors of Study Strategies, EUROA versus FIRSTAA*

Criterion variable	Predictor variable	$R^2$	Standardized $B$ weight	$F$ change	Sig.
Low anxiety	Speed	.10	0.32	16.60	.001
	Gender	.03	-0.16	4.24	.05
Select main ideas	Speed	.10	0.33	17.81	.01
	Separate knowing	.03	0.18	5.74	.05
	Success	.02	0.17	4.01	.05
	Speed	.13	0.37	23.15	.001
Testing strategies	Speed	.13	0.37	23.15	.001
	Separate knowing	.03	0.16	4.61	.05

Note. Sig. = significance.

Four factors predicted final course grade: Speed, Concentration, Gender, and Culture Contrast (EUROA vs. both Asian groups). Although beliefs about speed of learning, concentration on academic tasks, and gender contributed to final course grades, cultural differences remained.

Three factors predicted reading scores: Testing Strategies, English as First Language, and Study Aids. In sum, EUROA students differed from FIRSTAA students on reading scores. Cultural differences were eliminated once testing strategies, English as a first language, and use of study aids were entered into the regression.

### Discussion

The most fundamental question addressed in these analyses is: Are there differences in epistemological beliefs and ways of knowing among cultural groups? There were two significant differences. Euro-American students, compared to both Asian American groups, had stronger beliefs that learning is a slow, gradual process and that knowledge is complex. This is consistent with Qian and Pan's (2002) results, in which U.S. high school students believed more strongly that knowledge is complex than did high school students in China.

Additional analyses examined group differences in study strategies and academic performance. First, cultural differences in study strategies were revealed. Compared to first-generation Asian Americans, Euro-American students were better able to control their anxiety about school, select main ideas from texts, and prepare for tests. Compared to beyond-first-generation Asian Americans, European Americans were more able to select main ideas and use information processing strategies and were better academically motivated.

Next, we tested the notion that students' epistemological beliefs, ways of knowing, and gender may account for cultural differences

in study strategies. Results indicated that epistemological beliefs, ways of knowing, and gender contributed to cultural differences in study strategies.

Belief in quick learning contributed to cultural differences in study strategies. Students who scored lower on selecting main ideas, testing strategies, low anxiety, and high motivation were more likely to agree that learning is quick. For example, students who scored lower on these study strategies agreed with study strategy items such as the following: "When I am studying, worrying about doing poorly in a course interferes with my concentration" (Weinstein & Palmer, 2002, p. 9) and "When studying, I seem to get lost in the details and miss the important information" (Weinstein & Palmer, 2002, p. 11). These results suggest that if students believe learning is quick, they are likely to speed through homework and tests. Selecting main ideas would be difficult when rushing through a text. And it is easier to rush through easy parts of a task and skip the more difficult sections.

Belief in separate knowing and belief that knowledge is a constructive process also contributed to cultural differences in study strategies. Students who scored low on selecting main ideas, testing strategies, and information processing were less likely to believe in separate knowing. To understand the implications of these results, consider the students' scores for information processing. These scores involve students' use of strategies that lead to elaboration, organization, and linking prior knowledge with new knowledge (Weinstein & Palmer, 2002). The more students believe that knowledge is constructed and believe in separate knowing, the more likely they are to use information processing strategies. Indeed, if one believes knowledge is an integrated creation of ideas knitted together, one is more likely to engage in strategies that show connections between new ideas and one's own prior knowledge. Furthermore, engaging in evaluative or critical think-

Table 5

*Summary of Regression Analyses: Epistemological Beliefs as Predictors of Study Strategies, EUROA versus BEYONDAA*

Criterion variable	Predictor variable	$R^2$	Standardized $B$ weight	$F$ change	Sig.
Select main ideas	Speed	.09	0.29	14.53	.001
High motivation	Speed	.13	0.35	21.95	.001
	Culture contrast	.04	-0.21	7.37	.01
Information processing	Construction	.10	0.32	17.72	.001
	Separate knowing	.05	0.24	9.36	.01
	Culture contrast	.03	-0.18	6.20	.01

Note. Sig. = significance.

Table 6  
*Academic Performance Among Cultural Groups*

Variable	EUROA <i>M (SD)</i>	FIRSTAA <i>M (SD)</i>	BEYONDAA <i>M (SD)</i>
Course grade <sup>a</sup>			
Male	8.85 (3.27)	7.38 (3.30)	7.81 (3.22)
Female	10.03 (2.18)	8.44 (2.19)	8.04 (3.20)
Total	9.45 (2.81)	7.99 (2.77)	7.94 (3.19)
Reading score <sup>b</sup>			
Male	36.86 (3.34)	33.94 (5.74)	35.30 (5.23)
Female	36.32 (3.19)	34.40 (4.94)	35.02 (4.48)
Total	36.59 (3.26)	34.20 (5.28)	35.14 (4.80)

*Note.* Course grade ranged from A (12) to F (0). EUROA = Euro-Americans; FIRSTAA = first-generation Asian Americans; BEYONDAA = beyond-first-generation Asian Americans; Sig. = significance.

<sup>a</sup>EUROA was significantly different from both Asian groups.

<sup>b</sup>EUROA was significantly different from FIRSTAA.

ing would also encourage students to check consistency in logic between prior knowledge and new knowledge as well as consistency among the ideas presented.

Even more convincing evidence of the role of epistemological beliefs in learning was found in the examination of students' actual classroom performance for a business communication class. European Americans outperformed both Asian American groups in the business communication class. When epistemological beliefs, study strategies, and English as a native language were allowed to enter the equation to predict class performance, speed of learning and ability to concentrate predicted the final course grade.

Although Asian Americans are often perceived as outperforming Euro-American students, much of the past research that compared Asian American students to Euro-American students has focused on mathematics and science (Wang & Lin, 2005). When an overall math score was used as the dependent variable, Asian Americans typically outperformed Euro-American students. By contrast, when specific subtest scores were used, Asian Americans outperformed Euro-American students on calculation and routine problem solving. However, Euro-American students outperformed Asian American students on problems that were open-ended or required creativity.

It is important to keep in mind that the coursework in this study did not involve mathematics or science. Rather, the focus was on written and oral communication, group work, and critical thinking. For example, students in this study were assigned complex projects that required oral presentation and group exercises. Through these

assignments, students were required to demonstrate a proficiency in their verbal communication and business etiquette by planning, preparing, organizing, and effectively presenting various topics to multiple audiences. Specifically, this course work required creativity, cognitive flexibility, and a full array of skills.

This study provides impetus for future research in its limitations and questions left unanswered. First, this study did not specifically investigate the factor structure of epistemological beliefs or study strategies for Asian Americans. It is possible that these constructs do not manifest themselves in the same way as they do in European Americans, and this may have contributed to the somewhat lower interitem correlations (Cronbach's alpha coefficients) among some of the variables. Second, there was a significant age difference between EUROA students and BEYONDAA students. Thus, age is a potential confound in all comparisons between these two groups. This suggests that the next step is to conduct studies that carefully control for age and obtain multiple measures of epistemically related beliefs and study strategies. In addition, many questions remain unanswered. Would individual differences account for differences among cultures if the focus were on mathematics, history, computer science, engineering, or law? More importantly, if students were given instruction to enhance their study strategies and epistemically related beliefs (e.g., explicit instruction on knowledge integration, semester-long class projects, and study skills), would cultural differences decrease?

The present study supports the idea that epistemological beliefs are critical to learning and that cultural differences exist within learning environments. Furthermore, cultural stereotypes that denote one ethnic group as having the ideal or superior students should be avoided. These stereotypes could lead instructors to ignore students belonging to such groups who are struggling, yet reticent to ask for help (D. W. Sue & Sue, 2003; Wang & Lin, 2005). This stereotype could also be an important contributor to why students do not seek help from an instructor. For example, Lee (1994) writes about her ethnographic experiences with an academically struggling Asian American student:

When I suggested to Ming that he attend the tutoring sessions, Ming shook his head and said that he would not even consider attending these sessions. Ming said that it would be embarrassing to reveal his academic difficulties and that Asians did not talk about their problems. In his words: "You know, Asians don't talk about their problems—we just keep it inside—My father would kill me if I talk about stuff." Ming was referring to the Asian ethos that states that an individual's first loyalty is to his/her family and that "bad" behavior

Table 7  
*Summary of Regression Analyses: Epistemological Beliefs and Study Strategies as Predictors of Academic Performance*

Criterion variable	Predictor variable	<i>R</i> <sup>2</sup>	<i>b</i> weight	<i>F</i> change	Sig.
Final course grade	Speed	.09	0.21	18.90	.001
	Concentration	.05	0.11	10.85	.001
	Gender	.04	1.25	10.63	.001
	Culture contrast	.02	11.04	5.53	.05
Reading comprehension	Testing strategies	.11	0.30	17.39	.001
	Native English	.07	-2.57	11.44	.001
	Study aids	.03	0.15	4.59	.05

*Note.* Sig. = significance



(i.e., disclosure of failure) on the part of the individual shames the entire family (S. Sue & Sue, 1971). (Lee, 1994, p. 421)

For all intents and purposes, the avoidance of stereotyping is important for all students. Do not assume that a student slouched in his or her desk does not care about learning. Do not assume that a student who does not raise his or her hand to ask a question really has a strong understanding of the content. Do not assume that a student who dominates a group discussion is actually a strong leader in the class. Do not assume that a student who complains that there is too much homework is just lazy. Stereotypes lead to unchecked assumptions and subsequent default instructional techniques that can shortchange students.

The epistemically related beliefs of all students must be taken into consideration for instruction, testing, and intervention. Because cultural differences play an important role in learning, including research findings on epistemological beliefs and ways of knowing may enhance the classroom curriculum. Culturally sensitive curricula may alleviate student anxieties. It may help students to reassess their personal identities and self-esteem regarding their learning traits and practices. Such an approach to all learning environments, regardless of the content area, will provide a broader and pragmatic application of diversity research among college students.

## References

- Baxter Magolda, M. B. (1987). A comparison of open-ended interview and standardized instrument measures of intellectual development on the Perry scheme. *Journal of College Student Personnel*, 28, 443-448.
- Belenky, M. F., Clinchy, B. M., Goldberger, N. R., & Tarule, J. M. (1986). *Women's ways of knowing*. New York: Basic Books.
- Bendixen, L. D., & Rule, D. C. (2004). An integrative approach to personal epistemology: A guiding model. *Educational Psychologist*, 39, 69-80.
- Buehl, M. M., Alexander, P. A., & Murphy, P. K. (2002). Beliefs about schooled knowledge: Domain specific or domain general? *Contemporary Educational Psychology*, 27, 415-449.
- Cai, J. (1997). Beyond computation and correctness: Contributions of open-ended tasks in examining U.S. and Chinese students' mathematical performance. *Educational Measurement: Issues and Practice*, 16(1), 5-11.
- Cai, J. (1998). An investigation of U.S. and Chinese students' mathematical problem posing and problem solving. *Mathematics Education Research Journal*, 10(1), 37-50.
- Cai, J. (2000). Mathematical thinking involved in U.S. and Chinese students' solving of process-constrained and process-open problems. *Mathematical Thinking and Learning*, 2(4), 309-340.
- Chan, K. W., & Elliott, R. (2002). Exploratory study of Hong Kong teacher education students' epistemological beliefs: Cultural perspectives and implications on beliefs research. *Contemporary Educational Psychology*, 27, 392-414.
- Chang, D. F., & Demyan, A. (2007). Teachers' stereotypes of Asian, Black, and White students. *School Psychology Quarterly*, 22, 91-114.
- Clinchy, B. (2002). Revisiting *Women's Ways of Knowing*. In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 63-87). Mahwah, NJ: Erlbaum.
- Dweck, C. S., & Bempechat, J. (1983). Children's theories of intelligence: Consequences for learning. In S. G. Paris, G. M. Olson, & H. W. Stevenson (Eds.), *Learning and motivation in the classroom* (pp. 239-256). Hillsdale, NJ: Erlbaum.
- Easter, M., & Schommer-Aikins, M. (2004). Applying educational psychology to business communication: An initial model of cultural, epistemological, and relational views. *Communication Journal of New Zealand*, 5(1), 41-62.
- Galotti, K. M., Blythe, M. C., Ainsworth, K. H., Lavin, B., & Mansfield, A. F. (1999). A new way of assessing ways of knowing: The Attitudes Toward Thinking and Learning Survey (ATTLS). *Sex Roles*, 40(9/10), 745-766.
- Hofer, B. (2000). Dimensionality and disciplinary differences in personal epistemology. *Contemporary Educational Psychology*, 25, 378-405.
- Jehng, J. J., Johnson, S. D., & Anderson, R. C. (1993). Schooling and students' epistemological beliefs about learning. *Contemporary Educational Psychology*, 18(1), 23-35.
- Kao, G. (1995). Asian Americans as model minorities? A look at their academic performance. *American Journal of Education*, 103, 121-159.
- Karabenick, S. A., & Moosa, S. (2005). Culture and personal epistemology: U.S. and Middle Eastern students' beliefs about scientific knowledge and knowing. *Social Psychology of Education*, 8, 375-393.
- Kardash, C. M., & Wood, P. (2000, April). *An individual item factoring of epistemological beliefs as measured by self-reporting surveys*. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA.
- Kitchener, K. S., & King, P. M. (1981). Reflective judgment: Concepts of justification and their relationship to age and education. *Journal of Applied Developmental Psychology*, 2, 89-116.
- Kitchener, K. S., & King, P. M. (1989). The reflective judgment model: Ten years of research. In M. L. Commons, C. Armon, L. Kohlberg, F. A. Richards, T. A. Grotzer, & J. D. Sinnot (Eds.), *Adult development 2: Models and methods in the study of adolescent and adult thought* (pp. 63-78). New York: Praeger.
- Lee, S. J. (1994). Behind the model-minority stereotype: Voices of high- and low-achieving Asian American students. *Anthropology & Education Quarterly*, 25(4), 413-429.
- NCS Pearson. (2000). Reading and Arithmetic Indexes-12. Minneapolis, MN: Author.
- Ng, J. C., Lee, S. S., & Pak, Y. K. (2007). Contesting the model minority and perpetual foreigner stereotypes: A critical review of literature on Asian Americans in education. In L. Parker (Ed.), *Review of research in education: Difference, diversity, and distinctiveness in education and learning* (pp. 95-130). Los Angeles, CA: Sage.
- Perry, W. G., Jr. (1968). *Patterns of development in thought and values of students in a liberal arts college: A validation of a scheme* (ERIC Document Reproduction Service No. ED 024315). Cambridge, MA: Harvard University, Bureau of Study Counsel.
- Qian, G., & Alvermann, D. E. (2000). Relationship between epistemological beliefs and conceptual change learning. *Reading & Writing Quarterly*, 16, 59-74.
- Qian, G., & Pan, J. (2002). A comparison of epistemological beliefs and learning from science text between American and Chinese high school students. In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 365-385). Mahwah, NJ: Erlbaum.
- Rozin, P. (2003). Five potential principles for understanding cultural differences in relation to individual differences. *Journal of Research in Personality*, 37, 273-283.
- Ryan, M. P. (1984). Monitoring text comprehension: Individual differences in epistemological standards. *Journal of Educational Psychology*, 76, 248-258.
- Schoenfeld, A. H. (1983). Beyond the purely cognitive: Belief systems, social cognitions, and metacognitions as driving forces in intellectual performance. *Cognitive Science*, 7(4), 329-363.
- Schommer, M. (1990). The effects of beliefs about the nature of knowledge on comprehension. *Journal of Educational Psychology*, 82, 498-504.
- Schommer, M. (1994). Synthesizing epistemological belief research: Ten-

- tative understandings and provocative confusions. *Educational Psychology Review*, 6, 293–319.
- Schommer, M., Crouse, A., & Rhodes, N. (1992). Epistemological beliefs and mathematical text comprehension: Believing it is simple does not make it so. *Journal of Educational Psychology*, 84, 435–443.
- Schommer-Aikins, M. (2004). Explaining the epistemological belief system: Introducing the embedded systemic model and coordinated research approach. *Educational Psychologist*, 39, 19–29.
- Schommer-Aikins, M., Duell, O. K., & Hutter, R. (2005). Epistemological beliefs, mathematical problem-solving, and academic performance of middle school students. *Elementary School Journal*, 105(3), 289–304.
- Schreiber, J. B., & Shinn, D. (2003). Epistemological beliefs of community college students and their learning processes. *Community College Research and Practice*, 27(8), 699–710.
- Sue, S., & Sue, D. W. (1971). Chinese-American personality and mental health. *Amerasia Journal*, 1, 39–46.
- Sue, D. W., & Sue, D. (Eds.). (2003). *Counseling the culturally diverse: Theory and practice* (4th ed.). New York: Wiley.
- Tasaki, K. (2001). Culture and epistemology: An investigation of different patterns in epistemological beliefs across cultures. *Dissertation Abstracts International*, 62(2-A), 463. (UMI No. 3005226)
- Wang, J., & Lin, E. (2005). Comparative studies on U.S. and Chinese mathematics learning and the implications for standards-based mathematics teaching reform. *Educational Researcher*, 34(5), 3–13.
- Weinstein, C. E., & Palmer, D. R. (2002). *Learning and Study Strategies Inventory* (2nd ed.). Clearwater, FL: H&H.
- Wood, P., & Kardash, C. (2002). Critical elements in the design and analysis of studies of epistemology. In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge* (pp. 231–261). Mahwah, NJ: Erlbaum.

Received March 22, 2007

Revision received February 28, 2008

Accepted March 26, 2008 ■

### Members of Underrepresented Groups: Reviewers for Journal Manuscripts Wanted

If you are interested in reviewing manuscripts for APA journals, the APA Publications and Communications Board would like to invite your participation. Manuscript reviewers are vital to the publications process. As a reviewer, you would gain valuable experience in publishing. The P&C Board is particularly interested in encouraging members of underrepresented groups to participate more in this process.

If you are interested in reviewing manuscripts, please write to the address below. Please note the following important points:

- To be selected as a reviewer, you must have published articles in peer-reviewed journals. The experience of publishing provides a reviewer with the basis for preparing a thorough, objective review.
- To be selected, it is critical to be a regular reader of the five to six empirical journals that are most central to the area or journal for which you would like to review. Current knowledge of recently published research provides a reviewer with the knowledge base to evaluate a new submission within the context of existing research.
- To select the appropriate reviewers for each manuscript, the editor needs detailed information. Please include with your letter your vita. In the letter, please identify which APA journal(s) you are interested in, and describe your area of expertise. Be as specific as possible. For example, “social psychology” is not sufficient—you would need to specify “social cognition” or “attitude change” as well.
- Reviewing a manuscript takes time (1–4 hours per manuscript reviewed). If you are selected to review a manuscript, be prepared to invest the necessary time to evaluate the manuscript thoroughly.

Write to Journals Office, American Psychological Association, 750 First Street, NE, Washington, DC 20002-4242.



# Heuristics and Biases as Measures of Critical Thinking: Associations with Cognitive Ability and Thinking Dispositions

Richard F. West  
James Madison University

Maggie E. Toplak  
York University

Keith E. Stanovich  
University of Toronto

In this article, the authors argue that there are a range of effects usually studied within cognitive psychology that are legitimately thought of as aspects of critical thinking: the cognitive biases studied in the heuristics and biases literature. In a study of 793 student participants, the authors found that the ability to avoid these biases was moderately correlated with a more traditional laboratory measure of critical thinking—the ability to reason logically when logic conflicts with prior belief. The correlation between these two classes of critical thinking skills was not due to a joint connection with general cognitive ability because it remained statistically significant after the variance due to cognitive ability was partialled out. Measures of thinking dispositions (actively open-minded thinking and need for cognition) predicted unique variance in both classes of critical thinking skills after general cognitive ability had been controlled.

**Keywords:** critical thinking, heuristics and biases, intelligence, thinking dispositions

In the critical thinking literature, the ability to evaluate evidence and arguments independently of one's prior beliefs and opinions is a skill that is strongly emphasized (Baron, 1991, 2000; Ennis, 1987, 1996; Perkins, 1995; Sternberg, 1997, 2001, 2003). Virtually all measures of critical thinking try to assess the ability to avoid reasoning that is too biased by prior opinion and prior belief (e.g., Ennis, Millman, & Tomko, 1985; Facione, 1992; Norris & Ennis, 1989; Watson & Glaser, 1980).

The Watson-Glaser Critical Thinking Assessment (WGCTA; Watson & Glaser, 1980) is widely used and is representative of the discourse-logic-based critical thinking measures that have played a prominent role in critical thinking assessment. The WGCTA is an 80-item multiple choice test that comprises five subsets of items (Inference, Recognition of Assumptions, Deduction, Interpretation, and Evaluation of Arguments). Each test item generally consists of a series of statements about which the validity of various conclusions must be judged. Four of the WGCTA's five subscales emphasize the propositional logic of necessity rather than sufficiency. The test's authors acknowledge that strong opinions and beliefs have the potential to adversely influence people's ability to think critically, and the WGCTA includes both neutral

items and items on issues that are assumed likely to provoke a pressure to be biased. Thus, four of the five subscales admonish the test-taker to regard even factually questionable statements as correct for the purposes of the test (e.g., "For the purposes of this test, consider the statements in each exercise as true without exception"; "Try not to let your prejudices influence your judgment—just stick to the given statements [premises] and judge each conclusion as to whether it necessarily follows from the premises" [Harcourt Assessment, 2006, p. 4]). This emphasis on the importance of avoiding unbiased reasoning is also a common feature of the other discourse-logic-based critical thinking measures (e.g., Cornell Critical Thinking Tests, Ennis et al., 1985; California Critical Thinking Skills Test, Facione, 1992), and shares a common motivational underpinning with a number of tasks examined by cognitive scientists. Historically, Piaget's (1972) conceptualization of formal operational thought places such mechanisms of decontextualization—freeing from irrelevant context—in positions of paramount importance, because according to his view, "one of the essential characteristics of formal thought appears to us to be the independence of its form from reality content" (p. 10).

In the laboratory, the ability to reason in an unbiased manner has been operationalized with a few well-known paradigms. The belief bias syllogism task, one of the most thoroughly explored of these paradigms, shares a number of important features with discourse-logic-based critical thinking measures. This paradigm assesses the tendency for judgments of logical validity to be contaminated by prior knowledge of the world—for example, when the validity of a syllogism and the facts expressed in the conclusion of the syllogism conflict (e.g., "All flowers have petals; roses have petals; therefore, roses are flowers"—which is invalid under necessity). The inability to decouple prior knowledge from reasoning processes has been termed the *belief bias effect* (Evans, Barston, & Pollard, 1983). It has been the subject of extensive study in the cognitive science literature, and several formal models of how belief bias operates to disrupt syllogis-

---

Richard F. West, Department of Graduate Psychology, James Madison University; Maggie E. Toplak, Department of Psychology, Faculty of Health, York University, Toronto, Ontario, Canada; Keith E. Stanovich, Department of Human Development and Applied Psychology, University of Toronto, Toronto, Ontario, Canada.

This research was supported by a grant from the Social Sciences and Humanities Research Council of Canada to Maggie E. Toplak and by grants from the Social Sciences and Humanities Research Council of Canada and the Canada Research Chairs program to Keith E. Stanovich.

Correspondence concerning this article should be addressed to Richard F. West, Department of Graduate Psychology, MSC 7401, James Madison University, Harrisonburg, VA 22807. E-mail: westrf@jmu.edu



tic reasoning have been proposed (De Neys, 2006; Evans & Curtis-Holmes, 2005; Evans & Feeney, 2004; Garnham & Oakhill, 2005; Klauer, Musch, & Naumer, 2000).

Belief bias has also been revealed in paradigms in which participants must evaluate the quality of empirical evidence in a manner not contaminated by their prior opinion on the issue in question. In several studies, Klaczynski (1997) and colleagues (Klaczynski & Gordon, 1996; Klaczynski & Lavalley, 2005; Klaczynski & Robinson, 2000) presented participants with flawed hypothetical experiments that led to conclusions that were either consistent or inconsistent with prior positions and opinions. Participants then critiqued the flaws in the experiments (which were most often badly flawed). Participants found many more flaws when the experiment's conclusions were inconsistent with their prior opinions than when the experiment's conclusions were consistent with their prior opinions and beliefs.

It is, of course, important to show that the ability to reason independently of prior opinion is not entirely coexistent with general cognitive ability (intelligence), and there has been some preliminary evidence indicating that this is in fact the case. The tendency toward biased reasoning in the experiment evaluation paradigms (as well as related paradigms) shows considerable dissociation from cognitive ability (Kardash & Scholes, 1996; Klaczynski & Gordon, 1996; Klaczynski & Lavalley, 2005; Klaczynski & Robinson, 2000; Macpherson & Stanovich, 2007). Belief bias in syllogistic reasoning has shown a significant correlation with cognitive ability, but it is modest in size and does not exhaust the reliable variance in the magnitude of the bias displayed. We know the latter because various thinking dispositions (actively open-minded thinking, need for cognition) have been found to predict belief bias after the variance in cognitive ability has been partialled out (Kokis, Macpherson, Toplak, West, & Stanovich, 2002; Sá, West, & Stanovich, 1999; Stanovich & West, 1998).

The ability to reason independently of prior belief is only one component of critical thinking. Many theorists view critical thinking as a subspecies of rational thinking or at least as closely related to rational thinking (Kuhn, 2005; Moshman, 2004, 2005, in press; Reyna, 2004; Siegel, 1988, 1997). Cognitive scientists recognize two types of rationality: instrumental and epistemic. To think rationally means adopting appropriate goals, taking the appropriate action given one's goals and beliefs, and holding beliefs about the world that are commensurate with available evidence. These characteristics of rational thinking are precisely the features that a number of leading critical thinking theorists have highlighted in recent descriptions of critical thinking. Thus, Ennis (1996) describes critical thinking as "a process, the goal of which is to make reasonable decisions about what to believe and what to do" (p. xvii), Halpern (2008) states that "critical thinking is the use of those cognitive skills or strategies that increase the probability of desirable outcomes" (p. 3), and Facione (2007) emphasizes the importance of making purposeful, reflective judgments "about what to believe or what to do—precisely the kind of judgment which is the focus of critical thinking" (p.13).

If one accepts the theoretical linkage between critical thinking and rational thought, then there may well be other aspects of critical thought beyond the avoidance of egocentric processing that has been emphasized in previous work in critical thinking. However, we would also want any new domain of critical thinking to at least partially dissociate from measures of cognitive ability (as does the avoidance of belief bias). In this study, we examined one

candidate class of biases that qualify as indices of critical/rational thought and that may well dissociate from cognitive ability because tests of the latter do not assess it directly.

A prime candidate for a cognitive domain not assessed by tests of intelligence or tests of critical thinking is found in the panoply of effects studied in the heuristics and biases literature, which has a 30-year history in cognitive psychology (Evans, 1989, 2007; Evans & Over, 1996; Gilovich, Griffin, & Kahneman, 2002; Kahneman, 2003; Kahneman & Tversky, 1973, 1996, 2000; Over, 2004; Tversky & Kahneman, 1974, 1983, 1986). Many of the heuristics and biases studied relate to important aspects of rational and critical thought: causal reasoning, probabilistic reasoning, hypothetical thought, theory justification, assessment of the covariation of events, scientific reasoning, disjunctive reasoning, the tendency to think statistically, and the tendency to think of alternative explanations. These areas are legitimately classified as part of critical thinking. Research in the field of cognitive psychology has shown that these thinking characteristics can be measured and that they relate to important real-world decisions in domains such as personal finance, employment, health, and public policy (Baron, Bazerman, & Shonk, 2006; Hastie & Dawes, 2001; Hilton, 2003; Kahneman & Tversky, 2000; Lichtenstein & Slovic, 2006; Myers, 2002; Reyna & Farley, 2006; Reyna & Lloyd, 2006; Sunstein, 2002, 2005).

In short, on theoretical grounds, the traditional heuristics and biases studied by cognitive psychologists should be considered to be part of a broadened concept of critical thinking. We wish here to introduce theorists to this class of thinking skills because they are largely untapped by currently used critical thinking tests and to provide a preliminary indication of their likely empirical relationships to related constructs. In this investigation, we began some of the empirical groundwork for a theoretical integration of the heuristics and biases framework with more traditional approaches to critical thinking. We examined their relationship to an empirical marker of classic definitions of critical thinking (belief bias in logical reasoning) as well as their relationship to cognitive ability and thinking dispositions.

Although a heuristic process can result in behavior that is appropriate for a given purpose (e.g., unreflectively looking both ways before crossing a street), the set of tasks that we explored highlight important situations in which heuristics and biases are negatively associated with good critical thinking, and result in poor judgments and decisions about what to believe and what to do. Thus, in this article, we posit that the override of heuristics and the avoidance of biases are related to critical thinking.

We should emphasize that the large collection of skills and tasks from the heuristics and biases literature that we examined are not represented on traditional critical thinking tests like the WGCTA. For example, the WGCTA and similar genre of critical thinking tests do not assess the use of base rates in probabilistic reasoning; they do not tap knowledge of the law of large numbers or the ability of people to make regressive probabilistic predictions. They do not assess the gambler's fallacy, nor, in general, do they assess covariation detection, denominator neglect, or Bayesian probabilistic updating. In addition, they do not assess a critical aspect of rationality, *descriptive invariance*, that is the property that causes framing effects when violated (for reviews of the large body of literature on each of these effects, see Gilovich et al., 2002; Nickerson, 2004; Reyna, Lloyd, & Brainerd, 2003; Stanovich, 1999, 2004, 2008b). All of these effects and biases were examined in the present study.



In addition to examining the relationship between these varied measures of the ability to avoid heuristics and cognitive biases, we used the syllogistic reasoning task as an index of the type of discourse-logic-based reasoning skills that are assessed with traditional critical thinking tests like the WGCTA—the critical ability to reason logically when logic conflicts with prior belief. We examined whether cognitive ability was coextensive with variation on these two classes of critical thinking skills and whether cognitive ability mediated any connection between the two critical thinking skills. Additionally, we examined two thinking dispositions that have been found to be independent predictors of belief bias (need for cognition and actively open-minded thinking) in order to ascertain whether they were likewise independent predictors of the ability to avoid cognitive biases.

## Method

### *Participants and Procedure*

The participants were 793 undergraduate students (264 men and 529 women) recruited through an introductory psychology subject pool at a medium-sized state university. Their mean age was 19.0 years ( $SD = 2.1$ ). The majority of these students were freshmen (486 students) or sophomores (213 students), and a minority were juniors (65 students) and seniors (27 students). The year in college was missing for 2 participants. Almost 89% of them identified themselves as White (704 students), and a minority identified themselves as African American (20 students), Asian American (42 students), or “other” race/ethnicity (27 students).

Participants completed the battery of tasks during a single, 2-hr session. After completing the informed consent and demographic information sheet, participants completed, in order, the thinking disposition items, the syllogistic reasoning task, and the heuristics and biases tasks.

### *Tasks and Variables*

#### *Syllogistic Reasoning Problems with Belief Bias*

Twelve syllogistic reasoning problems, largely drawn from Markovits and Nantel (1989), were completed by the participants. Each problem was worded such that the validity judgment was in conflict with the believability of the conclusion. There were two types of these so-called inconsistent syllogisms. One type of inconsistent syllogism had a believable conclusion but an invalid format (e.g., “All flowers have petals; roses have petals; therefore, roses are flowers”—which is invalid). The other type had an unbelievable conclusions in a logically valid format (e.g., “All things with four legs are dangerous; poodles are not dangerous; therefore, poodles do not have four legs”—which is valid). Therefore, the believability of the content was inconsistent with the logical format of the syllogism in both types. Problems of this type have typically been thought to mirror the critical thinking skill of being able to put aside one’s prior knowledge and reason from new premises. After each item, the participants indicated their responses by selecting one of the two alternatives: (a) Conclusion follows logically from premises, or (b) Conclusion does not follow logically from premises. Scores on the inconsistent syllogisms ranged from 0 to 12 ( $M = 6.9$ ,  $SD = 3.0$ ).

#### *Heuristics and Biases Tasks*

*Causal Base Rate 1: Volvo problem.* In this problem, adapted from Fong, Krantz, and Nisbett (1986), a couple are deciding to

buy one of two otherwise equal cars. The *Consumer Reports* survey, statistics on repair records, and polls of experts favor the Volvo over the Saab. However, a friend reports experiencing a severe mechanical problem with the Volvo he owns. The participants were asked to provide advice to the couple. Preference for the Volvo indicates a tendency to rely on the large-sample information in spite of salient personal testimony. A preference for the Saab indicates reliance on the personal testimony over the opinion of experts and the large-sample information. Any degree of preference for the Volvo was scored as 1, and any degree of preference for the Saab was scored as 0.

*Causal Base Rate 2: Superintendent problem.* This is a slightly edited and revised version of a problem in Fong et al. (1986) that was analogous to the Volvo problem and was scored similarly. In this scenario, the superintendent of schools is urging the school board to make a curriculum shift. The superintendent must trade off the evidence of empirical studies conducted on a large sample of school districts that points in one direction versus the testimony of a school board member with a personal anecdote that points in the other direction.

*Noncausal base rate problem.* Base rates that have a causal relationship to the criterion behavior (Ajzen, 1977; Bar-Hillel, 1990; Tversky & Kahneman, 1979) are often distinguished from noncausal base rate problems—those involving base rates with no obvious causal relationship to the criterion behavior. In the previous two problems the statistical base rate information was causal (the *Consumer Reports* survey and the empirical studies of the school districts). In contrast, this problem is an adaptation of a noncausal base rate problem that has been much studied (Casscells, Schoenberger, & Graboys, 1978; Sloman, Over, Slovak, & Stibel, 2003; Stanovich & West, 1999, 2000). The problem was worded as follows:

It is known that 1 dollar out of every 10,000 is counterfeit. Imagine a money-changing machine that rejects real dollar bills 5 out of every 100 times it changes money. However, it always rejects bills when they are counterfeit. If this machine rejects your dollar bill, what is the probability (expressed as a percentage ranging from 0% to 100%) that your bill is counterfeit? Choose the best answer.

The problem was followed by the choices: (a) Less than 1%, (b) About 5%, (c) About 50%, (d) About 95%, and (e) More than 95%. Alternative a is the correct response and was scored as 1 (other responses were scored as 0).

*Law of large numbers: Hospital problem.* This problem is a classic and much-cited problem studied by Tversky and Kahneman (1974). It is used to explore participants’ understanding that, other things being equal, a larger sample size more accurately estimates a population value.

*Regression to the mean.* Drawn from Lehman, Lempert, and Nisbett (1988), this is a multiple-choice problem involving baseball batting averages, and only one of the alternatives shows some recognition of the possibility of regression effects and was scored as 1. Other options were scored as 0.

*Gambler’s Fallacy 1.* In the first gambler’s fallacy problem—the slot machine problem—the participant read the following:

When playing slot machines, people win something about 1 in every 10 times. Lori, however, has just won on her first three plays. What are her chances of winning the next time she plays? Choose the best answer.

The problem was followed by the choices: (a) She has better than 1 chance in 10 of winning on her next play, (b) She has less

than 1 chance in 10 of winning on her next play, (c) She has a 1 chance in 10 that she will win on her next play. The correct response of c was scored as 1, while any other response incorrect and scored as 0.

*Gambler's Fallacy 2.* This problem is similar to the last and involves coin tosses.

*Conjunction problem.* This problem is based on Tversky and Kahneman's (1983) much-studied Linda problem, one of the most studied problems in the heuristics and biases literature. Conjunction problems assess whether people appreciate that the probability of Event A and Event B both occurring must be lower than the probability of either A or B alone occurring.

*Covariation detection.* This problem appeared as follows:

A doctor had been working on a cure for a mysterious disease. Finally, he created a drug that he thought would cure people of the disease. Before he could begin to use it regularly, he had to test the drug. He selected 400 people at random who had the disease. Of the 400, he randomly assigned 300 to the treatment group and gave them the drug to see what happened. He randomly assigned 100 people to the no-treatment group and gave them a placebo (a sugar pill manufactured to look like the treatment drug) to see what happened. The table below indicates the outcome (Figure 1):

Drug condition	Group	
	Cured	Not cured
Received	200	100
Did not receive	75	25

Does the drug work?

Figure 1.

Participants were asked to choose the statement that best summarized the results shown in the table from among the following statements: (a) The evidence indicates that the drug was effective, (b) The evidence is inconclusive, or (c) The evidence indicates that the drug was not effective. The correct response, c, was scored as 1 (the other responses were scored as 0).

*Methodological reasoning.* Adapted from Lehman et al. (1988), this multiple-choice problem has only one alternative that indicates the ability to reason methodologically about confounded variables in everyday life. It was scored as 1 and the other responses as 0.

*Bayesian Reasoning 1.* This problem was the David Maxwell problem adapted from Beyth-Marom and Fischhoff (1983) and studied by Stanovich and West (1998). It is used to assess Bayesian belief updating.

*Bayesian Reasoning 2.* This problem was the Mark Smith problem adapted from Beyth-Marom and Fischhoff (1983) and studied by Stanovich and West (1998). It is used to assess Bayesian belief updating.

*Four card selection task.* Originally used by Wason (1966), the selection task has been studied extensively in the reasoning literature (e.g., Evans, Newstead, & Byrne, 1993; Evans & Over, 1996; Johnson-Laird, 1999). The problem involves reasoning about whether an "if-P-then-Q" type of rule can be falsified. Because the rule is in the form of an if-P-then-Q rule, the partic-

ipant must turn over the two cards that could potentially falsify the rule—the P card and the not-Q card (in this case, the Baltimore and train cards), which is the correct answer and was scored as 1. Very few participants give this completely correct response. Thus, we scored the task leniently by including as correct an alternative task construal championed by Margolis (1987). He has argued that turning the P card only is an appropriate response if the participant has adopted a so-called "open" reading of the rule—one in which the cards represent classes rather than individual exemplars. Therefore, the selection of the P card only was considered a correct answer and also scored as 1. All other selections were scored as 0.

*Framing problem.* The much-researched disease framing problem was originally studied by Tversky and Kahneman (1981). This problem is presented in two parts: positive and negative framing. Descriptive invariance is correct and was scored as 1. Violation of description invariance was scored as 0.

*Probabilistic reasoning: Denominator neglect.* This probabilistic reasoning task is a marble game that was modeled on a task introduced by Kirkpatrick and Epstein (1992; see also Denes-Raj & Epstein, 1994; Reyna, 1991; Reyna & Brainerd, 1994, 2007). The problem read as follows:

Assume that you are presented with two trays of black and white marbles: a large tray that contains 100 marbles and a small tray that contains 10 marbles. The marbles are spread in a single layer on each tray. You must draw out 1 marble (without peeking, of course) from either tray. If you draw a black marble, you win \$2. Consider a condition in which the small tray contains 1 black marble and 9 white marbles, and the large tray contains 8 black marbles and 92 white marbles. [A drawing of two trays with their corresponding numbers of marbles arranged neatly in 10-marbles-rows appeared above the latter sentence.] From which tray would you prefer to select a marble in a real situation?

The correct response is to select the small tray and was scored as 1.

*Disjunctive reasoning.* This problem was drawn from Levesque (1986, 1989) and is a variant on ones used by Toplak and Stanovich (2002). The problem was presented as follows:

There are 3 blocks in a stack, where each of the blocks is either new or old. The top block is new, and the bottom one is old. The middle block is either new or old. Is there a new block directly on top of an old block?

The text of the problem was accompanied by a graphic that displayed three blocks with their corresponding descriptions. The top block was labelled "new" the middle block was labelled "new or old", and the bottom block was labelled "old."

Participants were presented with the following alternatives: (a) Yes, (b) No, and (c) Cannot be determined. In order to solve this problem, one needs to consider the disjuncts of the middle box. That is, if the middle box is new, then the answer is "yes" because it would be on top of the bottom box, which is old. If the middle box is old, then the answer is still "yes" because the top box would now be on top of the middle box, which is old. The correct solution is thus "yes" and was scored as 1, and other responses were scored as 0.

*Heuristics and biases composite score.* For the purposes of several of the analyses reported below, we summed the scores on the 16 heuristics and biases tasks to form a composite score ( $M = 7.7$ ,  $SD = 2.6$ , range = 1–15). By forming a composite score, we do not mean to imply that these heuristics and biases tasks form a unidimensional construct. Previous intercorrelations of smaller sets of these tasks than those used here have shown that they are



only modestly correlated (Bruine de Bruin, Parker, & Fischhoff, 2007; Klaczynski, 2001; Parker & Fischhoff, 2005; Stanovich & West, 1998, 2000). The rational thinking tendencies measured by these heuristics and biases tasks are probably multifarious (Parker & Fischhoff, 2005; Reyna et al., 2003; Stanovich 2008a, 2008b; Toplak, Liu, Macpherson, Toneatto, & Stanovich, 2007). Finally, each task, from a psychometric point of view, represents only a single item. Thus, only modest reliability for the composite score is expected, and this was the case. The split-half reliability was .52, and Cronbach's alpha was .53.

### Cognitive Ability

Students were asked to indicate their verbal, mathematical, and total SAT scores on the demographics form. The mean reported verbal SAT score of the students was 584 ( $SD = 70$ ), the mean reported mathematical SAT score was 590 ( $SD = 68$ ), and mean total SAT score was 1,173 ( $SD = 102$ ). These self-reported scores closely matched the averages for this institution at the time of testing (582, 587, and 1,169, respectively). Several studies have indicated that the correlation between self-reported SAT scores and verified SAT scores is in the range of .80—.90 (Cassady, 2001; Kuncel, Crede, & Thomas, 2005) as is the correlation between self-reported grade point average (GPA) and verified GPA (Higgins, Peterson, Pihl, & Lee, 2007). An indication of the validity of the self-reported scores is that they correlated with a third variable to the same extent as verified scores. Stanovich and West (1998) found that the correlation between scores on a vocabulary test and self-reported SAT total scores (.49) was quite similar to the .51 correlation between the vocabulary test scores and verified total SAT scores in a previous investigation using the same vocabulary measure (West & Stanovich, 1991). The total SAT score is used as an index of cognitive ability in the analyses reported here because it loads highly on psychometric general intelligence as measured by a variety of indicators (Frey & Det-

terman, 2004; Unsworth & Engle, 2007). The total SAT score will be used in the analyses reported here.

### Thinking Dispositions

Two thinking dispositions were measured—actively open-minded thinking (Stanovich & West, 1997, 2007) and need for cognition (Cacioppo, Petty, Feinstein, & Jarvis, 1996). Items from the two scales were intermixed. We used the 41-item Actively Open-Minded Thinking (AOT) Scale (Stanovich and West, 2007). All items were scored such that higher scores represented a greater tendency toward open-minded thinking. Examples of items are “People should always take into consideration evidence that goes against their beliefs”; “Certain beliefs are just too important to abandon, no matter how good a case can be made against them” (reverse scored); and “No one can talk me out of something I know is right” (reverse scored). The responses for each item in the questionnaire were *strongly agree* (6), *moderately agree* (5), *slightly agree* (4), *slightly disagree* (3), *moderately disagree* (2), and *strongly disagree* (1). Higher scores on the scale indicate cognitive flexibility and lower scores indicate cognitive rigidity and resistance to belief change. The score on the scale was obtained by summing the 41 responses to the items ( $M = 170.5$ ,  $SD = 18.3$ ). The split-half reliability (Spearman-Brown corrected) of the scale was .76, and Cronbach's alpha was .84. The total score on the AOT scale was standardized, and the  $z$  score was used in statistical analyses.

The 18-item Need for Cognition Scale (Cacioppo et al., 1996) was used in this study. Sample items include “The notion of thinking abstractly is appealing to me,” and “I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.” The response format for each item in the questionnaire was the same as that used for the AOT. The mean score was 69.0 ( $SD = 12.2$ ). The split-half reliability (Spearman-Brown corrected) of the Need for Cognition Scale was .81, and Cronbach's alpha was .88. The total score on the scale was standardized, and the  $z$  score was used in statistical analyses.

Table 1

*Heuristics and Biases Tasks: Percentage Correct, Correlations, and Partial Correlations, With SAT Total Covaried*

Task	Percent correct	Correlation			Partial correlation
		SAT	Thinking dispositions <sup>a</sup>	Belief bias syllogisms	Syllogisms with SAT covaried <sup>b</sup>
Causal Base Rate 1	63.7	.211***	.128***	.215***	.139***
Causal Base Rate 2	77.3	.158***	.127***	.125***	.063
Noncausal base rate problem	47.9	.200***	.086*	.212***	.141***
Law of large numbers	41.7	.044	.057	.029	.010
Regression to the mean	39.6	.151***	.052	.155***	.099**
Gambler's Fallacy 1	50.9	.179***	.136***	.242***	.184***
Gambler's Fallacy 2	79.1	.115**	.104**	.091*	.045
Conjunction problem	27.5	.022	.033	.004	-.006
Covariation detection	31.7	.011	.063	.184***	.199***
Methodological reasoning	37.5	.062	.044	.066	.043
Bayesian Reasoning 1	61.8	.204***	.137***	.213***	.140***
Bayesian Reasoning 2	26.7	.237***	.178***	.274***	.195***
Four card selection task	19.9	.136***	.151***	.189***	.145***
Disease framing problem	62.2	.073*	.096**	.141***	.122***
Probabilistic reasoning: Denominator neglect	64.9	.191***	.092**	.149***	.073*
Disjunctive reasoning	37.7	.205***	.085*	.172***	.093**

Note.  $N = 793$ . SAT = SAT total.

<sup>a</sup> Thinking disposition composite. <sup>b</sup> Belief bias syllogisms.

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

A composite thinking dispositions score was also constructed. The  $z$  scores for the AOT and Need for Cognition scales were summed to create a composite score of these two measures. The reliability of the Composite Thinking Dispositions Scale was .88 (Cronbach's alpha).

### Results

Table 1 displays the percentage of participants who responded correctly on each of the heuristics and biases tasks. There is considerable variation in task difficulty. Consistent with the extensive literature on the difficulty of the four card selection task (e.g., Evans, 2006; Evans et al., 1993), this task was solved by less than 20% of the participants (the lowest solution rate of all the tasks). The easiest task was the second gambler's fallacy problem, which was answered correctly by almost 80% of the participants. More important is the fact that each of these tasks was answered incorrectly by at least substantial minorities. This is significant because, collectively, these tasks assess whether people adhere to some of the most fundamental strictures of rational thought (see Evans & Over, 1996; Gilovich et al., 2002; Kahneman & Tversky, 1996, 2000; Over, 2004; Samuels & Stich, 2004; Stanovich, 1999, 2004, 2008b; Stein, 1996). These results converge with those reported in a small body of work indicating that the susceptibility to these biases varies considerably (Bruine de Bruin et al., 2007; Klaczynski, 2001; Parker & Fischhoff, 2005; Stanovich & West, 1998, 2000, 2008). What predicts this variation in the susceptibility to these biases, and how does this variation relate to that in another foundational critical thinking skill—reasoning independently of prior belief? The next several analyses address these questions in various ways.

The second column in Table 1 lists the zero-order correlations between each of the 16 heuristics and biases tasks and the SAT total. Twelve of the 16 correlations in this column were significant. The four nonsignificant correlations were found for the law of large numbers, conjunction problem, covariation detection, and methodological reasoning problems. The third column lists the zero-order correlations between the heuristics and biases tasks and the thinking dispositions composite. Eleven of these 16 correlations were significant. Although the correlations between the reasoning tasks and thinking dispositions were more modest than those involving SAT, most once again reached a level of statistical significance. The five nonsignificant correlations were found for the law of large numbers, regression to the mean, the conjunction problem, covariation detection, and methodological reasoning problems. The fourth column lists the zero-order correlations between the heuristics and biases tasks and the belief bias syllogisms. Thirteen of these 16 correlations were significant. The three nonsignificant correlations were found for the law of large numbers, the conjunction problem, and methodological reasoning tasks. The final column displays the partial correlations between the heuristics and biases tasks and belief bias syllogisms after SAT was statistically covaried. Even after SAT was statistically partialled, 11 of the 16 correlations remained significant. Across the four columns displaying correlational data, only 3 of the 16 heuristics and biases tasks (law of large numbers, conjunction problem, and methodological reasoning) failed to demonstrate any reliable relationships with the other major measures.

Table 2 presents the zero-order correlations among the major variables in the study.<sup>1</sup> Because of the large sample size in the

Table 2

*Correlations Between Cognitive Ability, Thinking Dispositions, Heuristics and Biases Composite Score, and Syllogistic Reasoning With Belief Bias*

Variable	1	2	3	4	5	6
1. SAT total score	—					
2. AOT $z$ score	.205	—				
3. Need for Cognition $z$ score	.239	.401	—			
4. AOT and Need for Cognition composite	.265	.837	.837	—		
5. Heuristics and biases composite score	.390	.213	.247	.275	—	
6. Belief bias syllogisms	.441	.191	.225	.248	.437	—

*Note.* Values greater than  $r = .117$  significant at  $p < .001$ . AOT = Actively Open-Minded Scale.

study, all correlations were significant at the .001 level. The two components of critical thinking—avoidance of thinking biases and syllogistic reasoning independent of prior belief—displayed a moderate correlation with each other (.437). Both variables were moderately correlated with cognitive ability (.390 and .441, respectively) and modestly correlated with the composite thinking dispositions measure (.275 and .248, respectively). Because the association between the two thinking dispositions, the AOT and the Need for Cognition  $z$  scores, was relatively moderate ( $r = .401$ ,  $p < .001$ ), analyses paralleling those subsequently reported for the thinking disposition composite were also conducted on the AOT and on the Need for Cognition  $z$  scores separately.<sup>2</sup>

Previous research (Kokis et al., 2002; Stanovich & West, 1998) has found that variation in thinking dispositions related to cogni-

<sup>1</sup> One caveat concerning the associations that we observed in these studies relates to the restriction of range in our sample. Certainly, it is true that individuals with average and above-average cognitive ability are over-represented in samples composed entirely of university students. Nevertheless, the actual range in cognitive ability found among college students in the United States is quite large. In the past 30 years, the percentage of 25- to 29-years-olds in the United States who have attended college has increased by 50%. By 2002, 58% of these young adults had completed at least 1 or more years of college, and 29% had received at least a bachelor's degree (U.S. Department of Health and Human Services, 2003). However, the restriction of range in cognitive ability was somewhat greater in our sample, because our participants attended a moderately selective state university. The SAT total means of our sample were roughly .74 of a standard deviation above the national mean (College Board, 2006). The standard deviation of the distribution of scores in our sample was roughly .50 of the standard deviation in the nationally representative sample.

<sup>2</sup> Separate analyses involving the AOT and need for cognition thinking dispositions were also conducted in parallel to all of the analyses in this article in which the thinking dispositions composite score was used. Across the analyses, the pattern of findings for the two component dispositions closely matched those found for the composite measure. Both the AOT and need for cognition were individually significant predictors of performance in all of the regression analyses presented in Table 3, and both AOT and need for cognition accounted for comparable unique and common variance in the commonality analyses shown in Table 4. In each case, however, the thinking disposition composite measure accounted for a larger proportion of variance on the criterion variables.



tive flexibility predicts the avoidance of belief bias in reasoning after variation in cognitive ability has been controlled. The first regression analysis in Table 3 represents a replication of this finding. When entered first as a predictor of performance on the belief bias syllogisms, SAT total predicted 19.4% of the variance and, when entered second, the thinking dispositions composite predicted a statistically significant 1.9% of variance ( $p < .001$ ). The simultaneous regression of these two variables on belief bias syllogism performance indicated that cognitive ability was a more potent unique predictor (15.1% unique variance). A comparison of the simultaneous beta weights indicated the same thing.

The next regression analysis in Table 3 indicated that the thinking dispositions composite was likewise a unique predictor of the heuristics and biases composite score. When entered first as a predictor, SAT total predicted 15.2% of the variance and, when entered second, the thinking dispositions composite predicted a statistically significant 3.2% of variance ( $p < .001$ ). The simultaneous regression of these two variables on heuristics and biases performance indicated that cognitive ability was a more potent unique predictor (10.9% unique variance). A comparison of the simultaneous beta weights converged with this conclusion, although the variables were more balanced predictors than was the case in the analysis of syllogistic reasoning.

As shown in Table 2, the two components of critical thinking displayed a moderate correlation with each other (.437). The next analysis explored whether the link between the two was

entirely mediated by cognitive ability. The criterion variable in this analysis was the heuristics and biases composite score. When entered first as a predictor, SAT total predicted 15.2% of the variance and, when entered second, performance on the belief bias syllogisms predicted a substantial 8.7% of variance ( $p < .001$ ). The simultaneous regression of these two variables on the heuristics and biases score indicates that syllogistic reasoning performance was a more potent unique predictor than SAT total score (8.7% versus 4.8% unique variance). A comparison of the simultaneous beta weights provides a redundant way of comparing the potency of the two predictors.

The last regression in Table 3 examined whether the variance overlap between the susceptibility to various cognitive biases and the ability to reason independently of prior belief remained when the variance due to both cognitive ability and thinking dispositions was controlled. When entered first as a predictor of the heuristics and biases composite, SAT total predicted 15.2% of the variance. When entered second, the thinking dispositions composite predicted an additional 3.2% of variance ( $p < .001$ ). Finally, when entered last, performance on the belief bias syllogisms accounted for a substantial 7.4% additional variance. The simultaneous regression of these three variables on the heuristics and biases score indicated that syllogistic reasoning performance was a more potent unique predictor than either SAT total score or the thinking dispositions composite score (7.4% vs. 3.8% and 1.9% unique variance, respectively). A

Table 3  
*Hierarchical and Simultaneous Regression Results*

Criterion variable/order entered	Hierarchical		Simultaneous		
	$\Delta R^2$	$F$ to enter	$\beta$ weight	$F$ for $\beta$	Unique variance explained
Belief bias syllogisms					
1. SAT total	.194	190.67	.403	151.71	.151
2. Thinking disposition composite	.019	18.60	.141	18.60	.019
Overall regression					
$F = 106.76$					
Multiple $R = .461$					
Multiple $R^2 = .213$					
Heuristics and biases composite					
1. SAT total	.152	142.00	.341	104.80	.109
2. Thinking disposition composite	.032	30.53	.184	30.53	.032
Overall regression					
$F = 88.92$					
Multiple $R = .429$					
Multiple $R^2 = .184$					
Heuristics and biases composite					
1. SAT total	.152	142.00	.245	50.28	.048
2. Belief bias/syllogisms	.087	90.63	.329	90.63	.087
Overall regression					
$F = 124.36$					
Multiple $R = .489$					
Multiple $R^2 = .239$					
Heuristics and biases composite					
1. SAT total	.152	142.00	.218	39.32	.038
2. Thinking disposition composite	.032	30.53	.141	19.18	.019
3. Belief bias syllogisms	.074	78.38	.306	78.38	.074
Overall regression					
$F = 91.21$					
Multiple $R = .507$					
Multiple $R^2 = .258$					

Note. All  $F$  values in this analysis are significant at the  $p < .001$  level.

comparison of the simultaneous beta weights provides a redundant way of comparing the potency of the three predictors.

As an additional way to reveal the overlap in the variables as predictors of susceptibility to the cognitive biases, we conducted a commonality analysis (see Pedhazur, 1997) in which the variance explained by each variable was partitioned into a portion unique to that variable and portions shared with every possible combination of variables. The results of the commonality analysis are presented in Table 4. The first row indicates the unique variance in the heuristics and biases composite score explained by each of the predictors. The next row displays the explained variance in heuristics and biases scores that was common to SAT and thinking dispositions (.010). The third row displays the explained variance in heuristics and biases scores that was common to SAT and the belief bias syllogisms (.071). The fourth row displays the explained variance in heuristics and biases scores that was common to thinking dispositions and belief bias syllogisms (.013). The fifth row indicates that the explained variance in heuristics and biases scores that was common to all three predictors is .033. All of the variance components added together (.038 + .019 + .074 + .010 + .071 + .013 + .033) summed to the total variance explained in heuristics and biases scores by the three predictors (.258).

The unique variances explained here recapitulates those derived from the regression analyses. The commonality analysis reveals that the primary reason that SAT was not a more potent unique predictor was because of its overlap with variance explained by syllogistic reasoning performance (.071). Secondarily, the unique variance explained by cognitive ability (.038) was roughly similar to that in common to all three variables (.033). In contrast, syllogistic reasoning was more separable from other predictors. The proportion of unique variance that it accounted for was 2–3 times as large as that of the other two predictors, and it was the only indicator in which the amount of unique variance explained (.074) was larger than the explained variance held in common by all three predictors (.033).<sup>3</sup>

Table 4  
*Results of a Commonality Analysis Using the Heuristics and Biases Composite Score as a Criterion Variable*

Commonality	Variable		
	1. SAT total	2. Thinking dispositions <sup>a</sup>	3. Belief bias/syllogisms
Unique variance	.038	.019	.074
Common 1 & 2 <sup>b</sup>	.010	.010	—
Common 1 & 3 <sup>c</sup>	.071	—	.071
Common 2 & 3 <sup>d</sup>	—	.013	.013
Common 1 & 2 & 3 <sup>e</sup>	.033	.033	.033
Total unique variance + common variance	.152	.075	.191

*Note.* Unique variance = explained variance in heuristics and biases composite scores that is unique to that variable.

<sup>a</sup> Thinking disposition composite. <sup>b</sup> Common 1 & 2 = explained variance that was common to Variables 1 and 2. <sup>c</sup> Common 1 & 3 = explained variance that was common to Variables 1 and 3. <sup>d</sup> Common 2 & 3 = explained variance that was common to Variables 2 and 3. <sup>e</sup> Common 1 & 2 & 3 = explained variance that was common to all three variables.

## Discussion

The cognitive biases tapped by our heuristics and biases tasks are not typically assessed by measures of critical thinking. Yet there are theoretical reasons to believe that they should be. To perform well on these heuristics and biases tasks, one needs both declarative knowledge and the proper mental strategies and metastrategies (Stanovich, 2008a, 2008b). The heuristic response must be inhibited and replaced with a more normatively appropriate response. The inhibition process depends not only on cognitive capacity but also on the metacognitive sensitivity (i.e., reflectivity; see Stanovich, 2008b) to begin the process of response suppression. Once the heuristic response is inhibited, the knowledge to make the right substitute response must be present. Declarative knowledge of probability, causal reasoning, scientific thinking, and logic comes in to play at this step.

Several cognitive theorists have analyzed critical thinking in terms of rational thinking concepts and the philosophy of rational thought (e.g., Kuhn, 2005; Moshman, 2004, in press; Siegel, 1997). For example, discussions of critical thinking often concern the thought

<sup>3</sup> Structural equation modeling analyses provided additional evidence consistent with the pattern of findings reported in Table 3 and Table 4. Four different models were evaluated for relative fit. Model 1, which is most consistent with results displayed in the regression (Table 3) and commonality (Table 4) analyses, posited four separate latent constructs, with those of cognitive ability, thinking dispositions, and belief bias syllogistic reasoning simultaneously predicting the heuristics/biases endogenous construct. Three indicators were used for cognitive ability: total SAT score, high school grade point average (GPA), and college GPA (either high school or college GPA was missing for 78 participants). Two indicators were used for thinking dispositions: Need for Cognition Scale and AOT *z* scores. We created two item parcels each as separate indicators for the belief bias syllogism and heuristics/biases latent constructs by including odd-numbered and even-numbered items in each item parcel. This resulted in two indicators for both the belief bias syllogistic reasoning and the heuristics/biases composite.

Three additional models were tested to explore how well combining the latent constructs in various ways could account for the data. The following combinations were explored: when cognitive ability and thinking dispositions were treated as one of two exogenous constructs (Model 2); when cognitive ability and syllogistic reasoning were treated as one of two exogenous constructs (Model 3); and when cognitive ability, thinking dispositions, and syllogistic reasoning were treated as the one and only exogenous construct (Model 4). The heuristics and biases composite was the endogenous construct in each model. Model 1 generated  $\chi^2(21, N = 715) = 99.2, p = .0001$ ; confidence fit index (CFI) = .95; Tucker-Lewis index (TLI) = .91; root-mean-square error of approximation (RMSEA) = .07. Model 2 generated  $\chi^2(24, N = 715) = 176.2, p = .0001$ ; CFI = .90; TLI = .84; RMSEA = .09. Model 3 generated  $\chi^2(32, N = 715) = 208.6, p = .0001$ ; CFI = .88; TLI = .83; RMSEA = .09. Model 4 generated  $\chi^2(26, N = 715) = 303.2, p = .0001$ ; CFI = .81; TLI = .74; RMSEA = .12. While the chi-square statistic was significant for all of these models due to the large sample size in this study, Model 1 had the best fit of those tested on the basis of other indices used in this analysis. This suggests that considering these four constructs as separate latent variables is a good fit for this data.

The correlations and standardized weights from Model 1 also converged strongly with the results in Tables 3 and 4. Cognitive ability was significantly correlated with belief bias syllogisms ( $r = .57, p < .001$ ) and thinking dispositions ( $r = .44, p < .001$ ), and thinking dispositions were significantly correlated with syllogistic reasoning ( $r = .33, p < .001$ ). Further, the path coefficient from cognitive ability to heuristics and biases was significant (.32,  $p < .002$ ), the path from thinking dispositions to heuristics and biases was significant (.23,  $p < .001$ ), and the path from syllogistic reasoning to heuristics and biases was significant (.38,  $p < .0001$ ).



processes that we use to reason about what we should believe and how we should act given those beliefs. These thought processes correspond, respectively, to epistemic rationality (rationality of belief) and practical rationality (rationality of action). The distinction between epistemic and practical rationality has been much discussed in philosophy and cognitive science (Over, 2004; Samuels & Stich, 2004).

If these theorists are correct, then critical thought is rational thought—and the tasks used in our heuristics and biases battery tap some of the most important strictures of rational thinking that cognitive scientists have identified (Over, 2004; Samuels & Stich, 2004). For example, our battery examined violations of rational thinking such as conjunction errors in probability judgment, failures of descriptive invariance, ignoring sample size, failure to weight the denominator of the likelihood ratio, ignoring base rates, and failure to regress predictions (see Evans, 2007; Gilovich et al., 2002; Kahneman & Tversky, 1996, 2000; Stanovich, 1999, 2008b).

Our measure of the ability to avoid these thinking errors was moderately correlated ( $r = .437$ ) with a more traditional laboratory measure of critical thinking—the ability to reason logically when logic conflicts with prior belief. The correlation between these two classes of critical thinking skills was not due to a joint connection with cognitive ability. It remained statistically significant after the variance due to cognitive ability was partialled out (partial  $r = .321$ ,  $p < .001$ ). As the third regression in Table 3 shows, performance on the syllogistic reasoning task with belief bias predicted 8.7% unique variance ( $p < .001$ ) in performance on the heuristics and biases battery after cognitive ability was controlled, and as the fourth regression in Table 3 indicates, performance on the syllogistic reasoning task predicted 7.4% unique variance ( $p < .001$ ) in performance on the heuristics and biases battery after cognitive ability and thinking dispositions both were partialled out. As the commonality analysis in Table 4 indicates, the syllogistic reasoning task had as high a unique connection to heuristics and biases performance (7.4% variance explained) as it held in common with SAT as a predictor (7.1% variance explained).

Another triangulating conclusion that can be drawn from the commonality analysis in Table 4 is that our new class of critical thinking skills (the heuristics and biases tasks) is more specifically related to a classic critical thinking skill (the ability to reason logically when logic conflicts with prior belief) than it is to cognitive ability. The basis of this conclusion is that the unique variance in heuristics and biases performance explained by syllogistic reasoning performance (7.4%) was nearly double the unique variance explained by SAT scores (3.8%).

A notable finding of our study was that the thinking dispositions measures were independent predictors (after cognitive ability was controlled) of both classes of critical thinking skill (see the first and second regressions in Table 3). That thinking dispositions were an independent predictor of the ability to avoid bias by prior belief when reasoning converges with several previous studies. For example, Schommer (1990) found that a measure of the disposition to believe in certain knowledge predicted the tendency to draw one-sided conclusions from ambiguous evidence even after verbal ability was controlled. Kardash and Scholes (1996) found that the tendency to properly draw inconclusive inferences from mixed evidence was related to the disposition to believe in certain knowledge and to a measure of need for cognition. Furthermore, these relationships were not mediated by verbal ability because a vocabulary measure was essentially unrelated to evidence evaluation. Likewise, Klaczynski (1997; see also Klaczynski & Gordon, 1996; Klaczynski & Robinson, 2000) found that the degree to which

adolescents criticized belief-inconsistent evidence more than belief-consistent evidence was unrelated to cognitive ability (see also Perkins, Farady, & Bushey, 1991) but often was related to thinking dispositions associated with epistemic regulation. The evaluation of personally relevant arguments has been found to be related to actively open-minded thinking dispositions independent of cognitive ability (Sá et al., 1999; Stanovich & West, 1997).

The results of the first regression of Table 3—that our thinking dispositions measures predicted belief bias in syllogistic reasoning independent of cognitive ability—converges with these previous findings. What is even more notable is that the second regression in Table 3 indicates that the thinking dispositions were also unique predictors of a very different class of critical thinking skill—the susceptibility to the large class of cognitive biases discovered and reported in the heuristics and biases literature.

We wish here to introduce another class of thinking processes that needs to be considered when measuring the critical thinking abilities of individuals. Our goal has not been to develop a new critical thinking psychometric assessment instrument but, instead, to introduce theorists to a wide variety of critical thinking skills that are largely untapped by currently used critical thinking tests and to provide a preliminary indication of their likely empirical relationships to related constructs (cognitive ability and belief bias in reasoning). Typically, individual studies in the heuristics and biases literature have examined a single problem or only a few problems in an effort to understanding the processes that underlie responses that violate rational strictures (e.g., Gilovich et al., 2002; Kahneman, 2003; Kahneman & Tversky, 2000). Alternatively, the psychometric approach is characterized by the use of several similar items, which results in high internal consistency and reliability, as in measures of intelligence. In contrast, our approach has been to use a carefully selected set of different well-established heuristics and biases tasks that were expected to reflect relatively distinct cognitive skills, as opposed to a psychometric validation of a collection of problems thought to be characterized by similar reasoning tendencies. Thus, the rational thinking tendencies that we tapped with our measures reflect more than a single construct. Although some theoretical taxonomies have been proposed (Reyna et al., 2003; Stanovich 2008a, 2008b, in press; Stanovich, Toplak, & West, 2008), empirical work on construct differentiation is as yet virtually nonexistent—it is work we hope to provoke with our study. Given that we measured a multifarious concept with single tasks, the modest reliability of our heuristics and biases score was perhaps better than might have been expected. In the only studies (that we are aware of) in which multiple-item measures of different cognitive biases were used, Parker and Fischhoff (2005) and Bruine de Bruin et al. (2007) found mean interitem correlations of .12 and .16, respectively—higher than our mean of .066 but indicative that the biases are not reflective of a single underlying mechanism.<sup>4</sup> Further work will need to elaborate the structure of these biases so that they eventually might become more fully integrated with more classic notions of critical thinking which stress aspects of non-egocentric processing such as unbiased reasoning in the face of prior belief.

<sup>4</sup> An analysis of the correlation matrix in Table 2 of Klaczynski (2001) likewise yielded mean intertask correlations of roughly .15 among single-task measures of each bias when a single nonredundant variable was used for each task. Similarly, the mean correlation in a study including a battery of heuristics and biases tasks conducted by Slugoski, Shields, and Dawson (1993) was .03.



## References

- Ajzen, I. (1977). Intuitive theories of events and the effects of base-rate information on prediction. *Journal of Personality and Social Psychology*, 35, 303–314.
- Bar-Hillel, M. (1990). Back to base rates. In R. M. Hogarth (Eds.), *Insights into decision making: A tribute to Hillel J. Einhorn* (pp. 200–216). Chicago: University of Chicago Press.
- Baron, J. (1991). Beliefs about thinking. In J. Voss, D. Perkins, & J. Segal (Eds.), *Informal reasoning and education* (pp. 169–186). Hillsdale, NJ: Erlbaum.
- Baron, J. (2000). *Thinking and deciding* (3rd ed.). Cambridge, England: Cambridge University Press.
- Baron, J., Bazerman, M. H., & Shonk, K. (2006). Enlarging the societal pie through wise legislation. A psychological perspective. *Perspectives on Psychological Science*, 1, 123–132.
- Beyth-Marom, R., & Fischhoff, B. (1983). Diagnosticity and pseudodiagnosticity. *Journal of Personality and Social Psychology*, 45, 1185–1195.
- Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology*, 92, 938–956.
- Cacioppo, J. T., Petty, R. E., Feinstein, J., & Jarvis, W. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, 119, 197–253.
- Cassady, J. C. (2001). Self-reported GPA and SAT: A methodological note. *Practical Assessment, Research, and Evaluation*, 7(12).
- Casscells, W., Schoenberger, A., & Graboys, T. (1978). Interpretation by physicians of clinical laboratory results. *New England Journal of Medicine*, 299, 999–1001.
- College Board. (2006). *College Board SAT: 2006 college-bound seniors. Total group profile report*. Retrieved June 9, 2008, from [http://www.collegeboard.com/prod\\_downloads/about/news\\_info/cbsenior/yr2006/national-report.pdf](http://www.collegeboard.com/prod_downloads/about/news_info/cbsenior/yr2006/national-report.pdf)
- Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology*, 66, 819–829.
- De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science*, 17, 428–433.
- Ennis, R. H. (1987). A taxonomy of critical thinking dispositions and abilities. In J. Baron, & R. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 9–26). New York: Freeman.
- Ennis, R. H. (1996). *Critical thinking*. Upper Saddle River, NJ: Prentice-Hall.
- Ennis, R. H., Millman, J., & Tomko, T. N. (1985). *Cornell Critical Thinking Tests*. Pacific Grove, CA: Midwest.
- Evans, J. St. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hove, England: Erlbaum.
- Evans, J. St. B. T. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin and Review*, 13, 378–395.
- Evans, J. St. B. T. (2007). *Hypothetical thinking: Dual processes in reasoning and judgment*. New York: Psychology Press.
- Evans, J. St. B. T., Barston, J., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11, 295–306.
- Evans, J. St. B. T., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking & Reasoning*, 11, 382–389.
- Evans, J. St. B. T., & Feeney, A. (2004). The role of prior belief in reasoning. In J. P. Leighton & R. J. Sternberg (Eds.), *The nature of reasoning* (pp. 78–102). New York: Cambridge University Press.
- Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hove, England: Erlbaum.
- Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove, England: Psychology Press.
- Facione, P. (1992). *California Critical Thinking Skills Test & California Critical Thinking Dispositions Inventory*. La Cruz, CA: California Academic Press.
- Facione, P. (2007). *Critical thinking: What it is and why it counts*. Millbrae, CA: Insight Assessment, California Academic Press. Retrieved February 13, 2008, from [http://www.insightassessment.com/pdf\\_files/DEXadobe.PDF](http://www.insightassessment.com/pdf_files/DEXadobe.PDF)
- Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology*, 18, 253–292.
- Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or g? The relationship between the Scholastic Assessment Test and general cognitive ability. *Psychological Science*, 15, 373–378.
- Garnham, A., & Oakhill, J. V. (2005). Accounting for belief bias in a mental model framework: Comment on Klauer, Musch, and Naumer (2000). *Psychological Review*, 112, 509–518.
- Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. New York: Cambridge University Press.
- Halpern, D. (2008). *Halpern critical thinking assessment: Background and scoring standards*. Unpublished manuscript, Claremont McKenna College, CA.
- Harcourt Assessment. (2006). *Watson–Glaser Critical Thinking Appraisal: Sample questions* [Brochure]. Retrieved January 8, 2008, from <http://harcourt.assessment.com/hai/Images/dotCom/HTC/WG/SampleQuestions.pdf>
- Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world*. Thousand Oaks, CA: Sage.
- Higgins, D. M., Peterson, J. B., Pihl, R. O., & Lee, A. G. M. (2007). Prefrontal cognitive ability, intelligence, Big Five personality, and the prediction of advanced academic and workplace performance. *Journal of Personality and Social Psychology*, 93, 298–319.
- Hilton, D. J. (2003). Psychology and the financial markets: Applications to understanding and remedying irrational decision making. In I. Brocas & J. D. Carrillo (Eds.), *The psychology of economic decisions: Vol. 1. Rationality and well-being* (pp. 273–297). Oxford, England: Oxford University Press.
- Johnson-Laird, P. N. (1999). Deductive reasoning. *Annual Review of Psychology*, 50, 109–135.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80, 237–251.
- Kahneman, D., & Tversky, A. (1996). On the reality of cognitive illusions. *Psychological Review*, 103, 582–591.
- Kahneman, D., & Tversky, A. (Eds.) (2000). *Choices, value, and frames*. New York: Cambridge University Press.
- Kahneman, D. A. (2003). Perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, 58, 697–720.
- Kardash, C. M., & Scholes, R. J. (1996). Effects of pre-existing beliefs, epistemological beliefs, and need for cognition on interpretation of controversial issues. *Journal of Educational Psychology*, 88, 260–271.
- Kirkpatrick, L., & Epstein, S. (1992). Cognitive-experiential self-theory and subjective probability: Evidence for two conceptual systems. *Journal of Personality and Social Psychology*, 63, 534–544.
- Klaczynski, P. A. (1997). Bias in adolescents' everyday reasoning and its relationship with intellectual ability, personal theories, and self-serving motivation. *Developmental Psychology*, 33, 273–283.
- Klaczynski, P. A. (2001). Analytic and heuristic processing influences on adolescent reasoning and decision making. *Child Development*, 72, 844–861.
- Klaczynski, P. A., & Gordon, D. H. (1996). Self-serving influences on adolescents' evaluations of belief-relevant evidence. *Journal of Experimental Child Psychology*, 62, 317–339.
- Klaczynski, P. A., & Lavalley, K. L. (2005). Domain-specific identity,



- epistemic regulation, and intellectual ability as predictors of belief-based reasoning: A dual-process perspective. *Journal of Experimental Child Psychology*, 92, 1–24.
- Klaczynski, P. A., & Robinson, B. (2000). Personal theories, intellectual ability, and epistemological beliefs: Adult age differences in everyday reasoning tasks. *Psychology and Aging*, 15, 400–416.
- Klauer, K. C., Musch, J., & Naumer, B. (2000). On belief bias in syllogistic reasoning. *Psychological Review*, 107, 852–884.
- Kokis, J., Macpherson, R., Toplak, M., West, R. F., & Stanovich, K. E. (2002). Heuristic and analytic processing: Age trends and associations with cognitive ability and cognitive styles. *Journal of Experimental Child Psychology*, 83, 26–52.
- Kuhn, D. (2005). *Education for thinking*. Cambridge, MA: Harvard University Press.
- Kuncel, N. R., Crede, M., & Thomas, L. L. (2005). The validity of self-reported grade point averages, class ranks, and test scores: A meta-analysis and review of the literature. *Review of Educational Research*, 75, 63–82.
- Lehman, D. R., Lempert, R. O., & Nisbett, R. E. (1988). The effect of graduate training on reasoning. *American Psychologist*, 43, 431–442.
- Levesque, H. J. (1986). Making believers out of computers. *Artificial Intelligence*, 30, 81–108.
- Levesque, H. J. (1989). Logic and the complexity of reasoning. In R. H. Thomason (Ed.), *Philosophical logic and artificial intelligence* (p. 73–107). Dordrecht, the Netherlands: Kluwer Academic.
- Lichtenstein, S., & Slovic, P. (Eds.). (2006). *The construction of preference*. Cambridge, England: Cambridge University Press.
- Macpherson, R., & Stanovich, K. E. (2007). Cognitive ability, thinking dispositions, and instructional set as predictors of critical thinking. *Learning and Individual Differences*, 17, 115–127.
- Margolis, H. (1987). *Patterns, thinking, and cognition*. Chicago: University of Chicago Press.
- Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory & Cognition*, 17, 11–17.
- Moshman, D. (2004). From inference to reasoning: The construction of rationality. *Thinking and Reasoning*, 10, 221–239.
- Moshman, D. (2005). *Adolescent psychological development: Rationality, morality, and identity* (2nd ed.). Mahwah, NJ: Erlbaum.
- Moshman, D. (in press). The development of rationality. In H. Siegel (Ed.), *Oxford handbook of philosophy of education*. Oxford, England: Oxford University Press.
- Myers, D. G. (2002). *Intuition: Its powers and perils*. New Haven, CT: Yale University Press.
- Nickerson, R. S. (2004). *Cognition and chance: The psychology of probabilistic reasoning*. Mahwah, NJ: Erlbaum.
- Norris, S. P., & Ennis, R. H. (1989). *Evaluating critical thinking*. Pacific Grove, CA: Midwest.
- Over, D. E. (2004). Rationality and the normative/descriptive distinction. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 3–18). Malden, MA: Blackwell Publishing.
- Parker, A. M., & Fischhoff, B. (2005). Decision-making competence: External validation through an individual differences approach. *Journal of Behavioral Decision Making*, 18, 1–27.
- Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth, TX: Harcourt Brace.
- Perkins, D. N. (1995). *Outsmarting IQ: The emerging science of learnable intelligence*. New York: Free Press.
- Perkins, D. N., Farady, M., & Bushey, B. (1991). Everyday reasoning and the roots of intelligence. In J. Voss, D. Perkins, & J. Segal (Eds.), *Informal reasoning and education* (pp. 83–105). Hillsdale, NJ: Erlbaum.
- Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. *Human Development*, 15, 1–12.
- Reyna, V. F. (1991). Class inclusion, the conjunction fallacy, and other cognitive illusions. *Developmental Review*, 11, 317–336.
- Reyna, V. F. (2004). How people make decisions that involve risk. *Current Directions in Psychological Science*, 13, 60–66.
- Reyna, V. F., & Brainerd, C. J. (1994). The origins of probability judgment: A review of data and theories. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 239–272). New York: Wiley.
- Reyna, V. F., & Brainerd, C. J. (2007). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, 17, 147–159.
- Reyna, V. F., & Farley, F. (2006). Risk and rationality in adolescent decision making. *Psychological Science in the Public Interest*, 7, 1–44.
- Reyna, V. F., & Lloyd, F. J. (2006). Physician decision making and cardiac risk: Effects of knowledge, risk perception, risk tolerance, and fuzzy processing. *Journal of Experimental Psychology: Applied*, 12, 179–195.
- Reyna, V. F., Lloyd, F. J., & Brainerd, C. J. (2003). Memory, development, and rationality: An integrative theory of judgment and decision making. In S. L. Schneider & J. Shanteau (Eds.), *Emerging perspectives on judgment and decision research* (pp. 201–245). New York: Cambridge University Press.
- Sá, W., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology*, 91, 497–510.
- Samuels, R., & Stich, S. P. (2004). Rationality and psychology. In A. R. Mele & P. Rawling (Eds.), *The Oxford handbook of rationality* (pp. 279–300). Oxford, England: Oxford University Press.
- Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *Journal of Educational Psychology*, 82, 498–504.
- Siegel, H. (1988). *Educating reason: Rationality, critical thinking, and education*. London: Routledge.
- Siegel, H. (1997). *Rationality redeemed? Further dialogues on an educational ideal*. London: Routledge.
- Slooman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes*, 91, 296–309.
- Slugoski, B. R., Shields, H. A., & Dawson, K. A. (1993). Relation of conditional reasoning to heuristic processing. *Personality and Social Psychology Bulletin*, 19, 158–166.
- Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.
- Stanovich, K. E. (2004). *The robot's rebellion: Finding meaning in the age of Darwin*. Chicago: University of Chicago Press.
- Stanovich, K. E. (2008a). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond*. Oxford, England: Oxford University Press.
- Stanovich, K. E. (2008b). *The psychology of rational thought: What intelligence tests miss*. New Haven, CT: Yale University Press.
- Stanovich, K. E. (in press). *Rationality and the reflective mind: Toward a tri-process model of cognition*. New York: Oxford University Press.
- Stanovich, K. E., Toplak, M. E., & West, R. F. (2008). The development of rational thought: A taxonomy of heuristics and biases. *Advances in Child Development and Behavior*, 36, 251–285.
- Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, 89, 342–357.
- Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, 127, 161–188.
- Stanovich, K. E., & West, R. F. (1999). Discrepancies between normative and descriptive models of decision making and the understanding/acceptance principle. *Cognitive Psychology*, 38, 349–385.
- Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, 23, 645–726.
- Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning*, 13, 225–247.
- Stanovich, K. E., & West, R. F. (2008). On the relative independence of

- thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, 94, 672–695.
- Stein, E. (1996). *Without good reason: The rationality debate in philosophy and cognitive science*. Oxford, England: Oxford University Press.
- Sternberg, R. J. (1997). *Thinking styles*. Cambridge, England: Cambridge University Press.
- Sternberg, R. J. (2001). Why schools should teach for wisdom: The balance theory of wisdom in educational settings. *Educational Psychologist*, 36, 227–245.
- Sternberg, R. J. (2003). *Wisdom, intelligence, and creativity synthesized*. Cambridge, England: Cambridge University Press.
- Sunstein, C. R. (2002). *Risk and reason: Safety, law, and the environment*. Cambridge, England: Cambridge University Press.
- Sunstein, C. R. (2005). Moral heuristics. *Behavioral and Brain Sciences*, 28, 531–573.
- Toplak, M., Liu, E., Macpherson, R., Toneatto, T., & Stanovich, K. E. (2007). The reasoning skills and thinking dispositions of problem gamblers: A dual-process taxonomy. *Journal of Behavioral Decision Making*, 20, 103–124.
- Toplak, M., & Stanovich, K. E. (2002). The domain specificity and generality of disjunctive reasoning: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology*, 94, 197–209.
- Tversky, A., & Kahneman, D. (1974, September 27). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124–1131.
- Tversky, A., & Kahneman, D. (1979). Causal schemas in judgments under uncertainty. In M. Fishbein (Ed.), *Progress in social psychology*. Hillsdale, NJ: Erlbaum.
- Tversky, A., & Kahneman, D. (1981, January 30). The framing of decisions and the psychology of choice. *Science*, 211, 453–458.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293–315.
- Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business*, 59, 251–278.
- Unsworth, N., & Engle, R. W. (2007). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review*, 114, 104–132.
- U.S. Department of Health and Human Services, Office of the Assistant Secretary for Planning and Evaluation. (2003). Trends in the well-being of America's children and youth, 2003. Retrieved June 10, 2006, from <http://aspe.hhs.gov/HSP/03trends/index.htm>
- Wason, P. C. (1966). Reasoning. In B. Foss (Ed.), *New horizons in psychology* (pp. 135–151). Harmondsworth, England: Penguin.
- Watson, G., & Glaser, E. M. (1980). *Watson-Glaser Critical Thinking Appraisal*. New York: Psychological Corp.
- West, R. F., & Stanovich, K. E. (1991). The incidental acquisition of information from reading. *Psychological Science*, 2, 325–330.

Received September 25, 2007

Revision received May 6, 2008

Accepted May 8, 2008 ■

## Low Publication Prices for APA Members and Affiliates

**Keeping you up-to-date.** All APA Fellows, Members, Associates, and Student Affiliates receive—as part of their annual dues—subscriptions to the *American Psychologist* and *APA Monitor*. High School Teacher and International Affiliates receive subscriptions to the *APA Monitor*, and they may subscribe to the *American Psychologist* at a significantly reduced rate. In addition, all Members and Student Affiliates are eligible for savings of up to 60% (plus a journal credit) on all other APA journals, as well as significant discounts on subscriptions from cooperating societies and publishers (e.g., the American Association for Counseling and Development, Academic Press, and Human Sciences Press).

**Essential resources.** APA members and affiliates receive special rates for purchases of APA books, including the *Publication Manual of the American Psychological Association*, and on dozens of new topical books each year.

**Other benefits of membership.** Membership in APA also provides eligibility for competitive insurance plans, continuing education programs, reduced APA convention fees, and specialty divisions.

**More information.** Write to American Psychological Association, Membership Services, 750 First Street, NE, Washington, DC 20002-4242.



# Identifying Patterns of Appraising Tests in First-Year College Students: Implications for Anxiety and Emotion Regulation During Test Taking

Heather A. Davis  
North Carolina State University

Christine DiStefano  
University of South Carolina–Columbia

Paul A. Schutz  
University of Texas–San Antonio

The authors explored patterns of appraising tests in a large sample of 1st-year college students. Cluster analysis was used to identify homogeneous groups of 1st-year students who shared similar patterns of cognitive appraisals about testing. The authors internally validated findings with an independent sample from the same population of students and examined the extent to which cluster membership differentiated undergraduates on the basis of external indicators (e.g., anxiety, emotion-regulation strategies, and achievement). The authors used 2 randomly drawn samples to conduct an initial cluster analysis ( $n = 1,107$ ) and to replicate the solution on a 2nd, independent cluster and cross-classification analysis ( $n = 1,108$ ). There may be 5 subtypes of test takers who differ in how they approach tests, their experience of anxiety, and how they manage problems that occur during test taking. Theoretical implications for emotion and emotion regulation, as well as practical implications for working with undergraduates who experience test anxiety, are discussed.

**Keywords:** test taking, test anxiety, emotion regulation, cognitive appraisals, coping

Emotions, such as those experienced during testing, have been the subject of considerable theoretical and empirical work in the past 20 years (Ellsworth & Smith, 1988; Lazarus, 1999; Schutz & Davis, 2000; Schutz & Pekrun, 2007). This interest has emerged, in part, because of the roles that emotion and emotional regulation play in self-directed behavior. In terms of the research on emotions and emotional regulation, the area of test taking has received the most attention (Pekrun, Goetz, Titz, & Perry, 2002; Schutz & Pekrun 2007). The main focus of this research has been on the nature of test anxiety (e.g., Hodapp & Benson, 1997; Sarason & Sarason, 1990; Zeidner, 1998, 2007) and on how students cope or regulate their anxiety during the testing process (Carver & Scheier, 2000; Folkman & Lazarus, 1985; Kondo, 1997; Zeidner, 1998). This research has contributed to the understanding of emotion and emotional regulation during test taking and to the development of a context-specific, valid, and reliable measure of emotional regulation during test taking (Schutz, Benson, & DeCuir, in press; Schutz, DiStefano, Benson, & Davis, 2004). However, to date, there have been few studies (Davis, DiStefano, DeCuir, & Schutz, 2000; see also Brdar, Rijavec, & Loncaric, 2006; DeCuir, Aultman, & Schutz, in press; Rice & Slaney, 2002; Sondaite &

Zukauskienė, 2005; Tanaka, 2007) that have attempted to systematically identify whether there are group differences in the ways students approach academic tasks—specifically tests—and in the strategies they deploy to attempt to manage their emotions.

In his study of 100 high school and 241 college students, Zeidner (1996; see also Zeidner & Matthews, 2005) found that students with richer coping resources concurrently tended to report lower levels of trait anxiety and evidenced lower levels of state anxiety during an exam. Zeidner argued that testing situations contain the critical elements of major environmental stressors, including the need to prepare for an impending event, eminent confrontation with a stressor, uncertainty surrounding the outcome of the event, and the need to cope with the consequences. Students were asked to report their level of test anxiety, their coping resources, and their situational coping strategies 6 weeks prior to an exam. Dispositional judgments were then used to predict their level of state anxiety during an exam. Zeidner found that there was no relationship between the deployment of coping strategies and the resultant experience of anxiety during tests.

In the following sections, we begin by defining what emotion regulation is and how the judgments students make about test taking may contribute to a “proneness” (Zeidner, 2007, p. 167) to experience anxiety. Next, we review four theories of emotion processes as they relate to achievement activities. Each of these theories makes specific claims about which judgments, or appraisals, are the most salient in the study of achievement tasks. These theories describe the ways in which students’ perceptions ultimately affect the quality of their engagement and performance. Finally, we describe a project designed to examine the ways in which college students may approach tests in systematically different ways—and the implications of these findings for theories of

---

Heather A. Davis, Department of Curriculum and Instruction, North Carolina State University; Christine DiStefano, Department of Educational Studies, University of South Carolina–Columbia; Paul A. Schutz, Department of Educational Psychology, University of Texas–San Antonio.

We acknowledge the invaluable contributions of Claire Ellen Weinstein, who included us on this project and provided guidance along the way.

Correspondence concerning this article should be addressed to Heather A. Davis, 402J Poe Hall, Box 7801, Raleigh, NC 27695-7801. E-mail: Heather\_Davis@ncsu.edu

emotion regulation and for working with populations of test anxious students.

### Understanding Evaluation Anxiety and Emotion Regulation

"Evaluation anxiety can have serious consequences for one's physical and mental health as well as for one's educational achievement and occupational career" (Zeidner, 2007, p. 165). In a recent meta-analysis, Hembree (1998) found test anxiety to be negatively correlated with achievement in college and high school populations as well as being related to performance and well-being. According to Zeidner (2007), "test anxiety has been found to interfere with competence in both lab settings as well as in true-to-life test taking situations in school and college" (p. 171).

Scholars who study test anxiety tend to make at least two important distinctions. The first is to distinguish the difference between trait anxiety and state anxiety (Zeidner, 1995). The second is to distinguish between the cognitive and affective components of the anxiety experience. Zeidner (2007) described trait anxiety as reflecting individual differences in students' "prone-ness" (p. 167) toward feeling anxious during a test, with some students experiencing pervasive or excessive worry about exams even when they are not in the immediate testing situation. "Trait like affect reflects a general way of responding to the world, which varies by person, but is relatively stable over time" (Linnenbrink, 2007, p. 108; see also Rosenberg, 1998). In contrast, state test anxiety may fluctuate depending on the context of a given test and can be brought about by changes in the testing environment.

In the short-term, acute distress and worry are generated by accessing negative self-beliefs, that one lacks academic competence, for example, and by choosing counterproductive coping strategies, such as self-blame and avoidance, that focus attention on personal shortcomings in the academic domain. (Zeidner, 2007, p. 167)

In distinguishing between the cognitive and affective components of test anxiety, scholars have attempted to tease out the thoughts and beliefs that lead students to perceive threats in the testing context as somewhat distinct from the different forms of arousal (e.g., tension, bodily reactions) they may feel while taking tests. Our focus is on these self-referent thoughts and beliefs. In fact, cognitive appraisal theory suggests that the judgments students make about their testing situation serve as organizers for their affective response system (van Reekum et al., 2004).

### *Cognitive Appraisals About Tests*<sup>1</sup>

Research on cognitive appraisals has suggested that the judgments, or appraisals, people make about their environment and their ability to manage challenges/threats are related to the emotions they experience (Lazarus, 1991; Roseman & Smith, 2001; Smith, 1991). These judgments can occur rapidly and without conscious awareness but are viewed as essential for emotions to emerge (Panksepp, 2005). Appraisal theory asserts three important assumptions. First, cognitive appraisal theory assumes the appraisal process takes place continuously. Because of this, one way to regulate or manage an emotion may be to reappraise the situation, that is, to see the situation in a new light. Folkman and Lazarus (1985) published a seminal series study in which they

assessed students' emotions and coping at three phases of the examination process: anticipation of the exam, the wait for grades to be posted, and the time after grades were posted. At each stage, they were asked to describe their state of mind and rate their experience of emotions associated with perceiving threat (i.e., worry, fear, and anxiety), challenge (i.e., confidence, hope, and eagerness), harm (i.e., anger, sadness, disappointment, and guilt) and benefit (i.e., exhilaration, pleasure, happiness, and relief). They found there were significant changes in the emotions students experienced, including seemingly contradictory emotions, as well as in the coping strategies they deployed. These findings suggested that the appraisal process was ongoing, throughout the entire examination process, resulting in a fluctuating emotion experience.

Second, appraisal theory asserts that the judgments are relational to the individual with "goals provid[ing] the 'direction' in self-regulation" (Schutz & Davis, 2000, p. 245). That is why two students can look at the same test, can both perceive it as challenging, but only one may become anxious. People's cumulative histories and their beliefs about themselves contribute to the in-the-moment judgments they make and the emotions that result. In an interesting series of studies by Kondo (1997), 79 undergraduate students were asked to complete a measure of test anxiety and then to describe the tactics they used to cope with their anxiety in anticipated and actual examinations. Kondo generated a cumulative list of all strategies and then had a second set of undergraduates sort these strategies into meaningful categories. He then correlated the types of strategies deployed with the original 79 participants' anxiety scores. In addition to finding that some types of test preparation strategies were more closely associated with the experience of anxiety, Kondo found that "people with high test anxiety tended to report significantly more strategies than those with low test anxiety" (p. 210). Thus, consistent with appraisal theory, some students who experience higher test anxiety may do so as a function of holding higher standards for performance.

In the appraisal literature, Lazarus (1991, 1999) made a distinction between primary and secondary appraisals. Primary appraisals are associated with the likelihood and intensity of the emotion experience, whereas secondary appraisals are associated with flavoring the specific emotion experienced. For primary appraisals, there are two pivotal judgments, labeled *goal relevance* and *goal congruence*, which are viewed as defining the likelihood of an emotion to emerge. Goal relevance is defined as the judgments individuals make about how important events are to their goals. In other words, for test anxiety to emerge, students must judge tests, and test scores, as important to achieving their academic or personal goals. If tests are not seen as important, the potential for feeling anxiety diminishes. Goal congruence is defined as the extent to which events are perceived as aligning with individuals' goals. Events that are perceived to be in conflict with or obstacles to goal pursuit are goal incongruent. Within the testing context, the appraisal of goal congruence is related to students' perceptions of how helpful a particular test or test scores in general are in moving them closer to their goals. These goals might include acquiring a

<sup>1</sup> We thank the reviewers for their invaluable feedback about the nature of individual differences in appraising tests. Their feedback led to major revisions of the content and structure of the literature review.



specific grade in a class, graduating, or preparing for a specific career. According to appraisal theory, pleasant emotions, such as happiness or pride, are more likely to be experienced during tests that are perceived to be goal congruent, whereas unpleasant emotions, such as anxiety, anger, or shame, are more likely to emerge when tests are perceived as goal incongruent. Again, it is important to keep in mind that students use their own standards when judging whether a test is going well or poorly. Doing well for one student might involve understanding some but not all items on the test, whereas for other students, having to stop and think, or reason through, any single item on a test may feel threatening because they are using a standard of perfection or an A as their goal (Schutz & Davis, 2000).

Third, appraisal theory assumes that the secondary appraisal process is differentiated, with some judgments mattering more than others in bringing about any given emotion (Smith, Haynes, Lazarus, & Pope, 1993). Several different programs of research have attempted to develop taxonomies, or classification systems, to predict which appraisals lead to different emotion experiences (Roseman, Anoniou, & Jose, 1996; Scherer, 1988; Smith, 1991). Ben-Ze'ev (2000) provided one of the most comprehensive descriptions of the numerous different types of appraisals that have been studied to predict and describe emotion experiences.

Within the field of test taking, Schutz and Davis (2000; see also Schutz et al., 2004, in press; Smith & Ellsworth, 1987) argued that there are two key secondary appraisals involved in eliciting test anxiety: agency and testing problem efficacy. *Agency* is defined as the extent to which students appraise the outcome of a particular test as being under their control. According to Schutz and colleagues (Schutz & Davis, 2000; Schutz et al., 2004, in press), for anxiety to emerge, students must perceive a lost sense of agency related to the test. Kondo (1997) suggested that the perception of lack of control may reflect the uncertainty students feel about the extent to which their preparation efforts matched or did not match the demands of the test. In addition, *testing problem efficacy* is defined as the judgments students make about their ability to manage the problems that emerge during tests, such as not understanding a question or having to reason an answer through. They argued that anxiety is more likely to emerge when students judge they have little or no confidence in their ability to manage the problems or their emotions during tests.

Fourth, Pekrun, Frenzel, Goetz, and Perry (2007) discussed the ways in which emotions may become more "habitualized" (p. 23) as a function of repeated associations with a given situation. As emotion experiences become associated with tasks, the initial appraisal process may be short-circuited, and the appraisal processes become more consistent. In other words, when faced with a new test, students not only evaluate the demands of a specific test, but also draw on their past experiences taking tests and their past judgments of goal relevance, congruence, agency, and test problem efficacy. These patterned ways of appraising tests can help researchers understand how discrete judgments of tests can, over time, affect more general outcomes, such as achievement across courses or on standardized measures. In fact, Schutz et al. (2004) found that appraisal processes contributed to predicting between 16% and 25%, or between a half and a full letter grade, of the variance in students' self-reported college grade point average.

### *Defining Emotion Regulation and Regulating During Tests*

Emotion regulation has been defined as the physiological, behavioral, and cognitive processes that enable individuals to modulate their experiences and expression of emotions (Gross, 1999; Gross & Munoz, 1995; Gross & Thompson, 2007; John & Gross, 2004). In some cases, moderating may mean dampening, or down-regulating, the experience of an emotion, whereas in other cases, moderating may involve amplifying, or up-regulating, an emotion. Gross and Thompson (2007) noted that emotion regulation occurs on a continuum from conscious, effortful, and controlled to unconscious, effortless, and automatic. Also, it is important to understand that merely engaging in some form of regulation (i.e., deploying some strategy) can make something better or worse depending on whether the selection was appropriate or enactment was effective. Restated within the context of test taking, although some students may consciously engage in trying to reduce unpleasant feelings during tests, their enactment of specific strategies may not necessarily produce the results they desire. This is what makes studying emotion regulation so challenging. A strategy that may be defined as less adaptive by the field might actually serve a functional purpose for a given student, and strategies that have been historically defined as adaptive may not assist students in modulating their emotion experience if they are enacted poorly (John & Gross, 2007; Lazarus, 1985; Thompson & Calkins, 1996). Thompson and Calkins (1996) argued that there may be a cost-benefit trade-off for engaging in strategies aimed at preserving or changing one's sense of self. That is, strategies deployed to dampen an unpleasant emotion may provide students with a short-term benefit; however, such strategies may have the hidden consequence of diminished performance.

Gross and Munoz (1995; see also Gross & Thompson, 2007) outlined five different families of emotion-regulation strategies: situation selection, situation modification, attentional deployment, cognitive change, and response modulation. In situation selection, individuals avoid certain situations, people, or contexts to try to decrease unpleasant emotions or, in turn, seek out specific situations, people, or contexts as a way of enhancing pleasant emotions. In situation modification, people seek to assert themselves into a situation to modify that situation to meet their needs. Within the testing context, most students do not have the opportunity to engage in this type of regulation. Tests, in large part, are determined by instructors—although some instructors may be willing to negotiate format, timing, and so forth. In attentional deployment, individuals try to focus on more or less valued aspects of the context to up- or down-regulate an emotion. The goal is either to distract themselves from or to concentrate on elements of the situation. Davis, Schutz, and DeCuir (1999) found that students with lower levels of test anxiety used attentional deployment to regulate their experience of anxiety by concentrating first on the easy questions on tests and then returning to work on harder questions. In many cases, this helped to reduce the students' anxiety because they could compare the numbers of easy and hard questions, and they felt confident about their understanding of the material (i.e., given their responses to easy questions) by the time they began working on more challenging questions. In cognitive change, individuals try to construct a more positive meaning out of their experience, for example, by reappraising the relative impor-



tance or value of the situation. Each of these four types of regulation strategies could be considered antecedent-focused strategies, in that, when deployed, their intent is to try to alter some element of the context, beliefs, or judgments that bring about the emotion experience. In contrast, in response modulation, individuals attempt to work on the emotion experience by trying to suppress or extract the emotion they feel. Although many scholars have studied each of these strategies, both in lab and naturalistic settings, Gross and Thompson (2007) noted, "One important and as yet unanswered question is how different forms of emotion regulation typically co-occur" (p. 17).

Within the field of emotion regulation, the largest body of literature exists in the area of coping (Lazarus & Folkman, 1984, 1991). Lazarus (2001) defined coping as "the effort to manage psychological stress" (p. 45). Early work in the field of coping identified two different forms of coping—problem focused and emotion focused (Carver, Scheier, & Weintraub, 1989). In problem-focused coping, individuals deploy strategies to manage the demands of the task, whereas in emotion-focused coping individuals deploy strategies to modulate the emotion experience. According to several studies aimed at developing measures to capture coping, some individuals tend to deploy, and even rely on, certain strategies in certain situations (Carver & Scheier, 2000). Moreover, a large body of research has indicated that there may be consequences for reliance on emotion-focused coping strategies, such as higher reports of depression (Rafnsson, Jonsson, & Windle, 2006) and unpleasant emotions like anger (Martin & Dahlen, 2005). In fact, most researchers in the field of coping have suggested that problem-focused coping strategies are generally more productive or adaptive (Moos & Holahan, 2003).

Within the field of test taking, Schutz and colleagues (Schutz & Davis, 2001; Schutz et al., 2004, in press) have identified three different dimensions of coping that students use to manage problems during tests: task-focused processes, regaining task focus, and emotion-focused processes. The key element of task-focused processes is that the focus of students' internal talk is the strategies they are using during the task. For example, during testing, students' task-focused thoughts might be about how to manage their time during the test or about looking for answers to one question in another. These thoughts help to keep students focused on the test and away from potentially disruptive thoughts about themselves. In keeping with the work of Gross and colleagues (Gross, 1998, 2002; Gross & Munoz, 1995; Gross & Thompson, 2007), task-focused processes largely reflect students' attempts at attentional deployment. In this case, students focus on those elements of the test they can control: reading the directions, finding main ideas in the questions, eliminating responses, and so forth. This shift in concentration away from what confuses them and onto what they understand not only regulates the emotion, but also manages the actual demands of the test.

Aligned with earlier research on coping, the second dimension of emotion-focused processes involves a shift in students' focus from the task to themselves and the emotions related to the task. It involves a disengagement from the task and a focus on their feelings and thoughts about their performance on the task and the potential causes for that performance. These emotion-focused thoughts tend to distract the examinee from the task and may lower achievement. In keeping with the work of Gross and colleagues (Gross, 1998, 2002; Gross & Munoz, 1995; Gross & Thompson,

2007), emotion-focused coping strategies also largely appear to reflect attempts at attentional deployment (to focus on how teachers may score the test or events outside of the testing context) or cognitive change (to change the meaning of challenges on a test by assuming responsibility for not preparing).

Last, Schutz et al. (2004, in press) argued that a third dimension exists, called regaining-task-focus processes. This dimension comprises processes that involve students' attempts to get back on task by attempting to reduce their tension or put the test in perspective. As such, these processes tend to be neither completely task focused nor completely emotion focused. In keeping with the work of Gross and colleagues (Gross, 1998, 2002; Gross & Thompson, 2007), these two strategies appear to target different families, with tension reduction reflecting students' attempts at response modulation (i.e., suppressing unpleasant feelings) and importance reappraisal reflecting students' attempts at cognitive change (i.e., reconstructing the meaning of the test in their lives).

### Theorizing Systematic Differences in Emotional Regulation During Testing

The purpose of this project was to explore whether students evidence habitual patterns of appraising tests that might make them more prone to experiencing anxiety during tests. In the following sections, we briefly review three theories of emotionality in education that have important implications for understanding why anxiety might emerge during a test, how the experience of anxiety during a test might be interpreted, and how students might systematically differ in their approaches toward tests and their attempts to modulate the experience of test anxiety.

#### *Circumplex Model of Affect*

In her circumplex theory of emotions in education, Linnenbrink (2007), building on the work of Barrett (2006) and Russell and Barrett (1999), described the relationship between emotions and motivation as a function of valence, what we have thus far been referring to as goal congruence, and activation. She defined activation as the experience of "arousal, mobilization, and energy" (p. 108). That is, emotions that are activating have a directional component that can compel students toward or away from different activities, tasks, or behaviors. Furthermore, Linnenbrink described the ways emotions affect people's behavioral and cognitive engagement depending on whether they are pleasant or unpleasant (i.e., valence) or activating/deactivating. *Behavioral engagement* is defined as the amount and quantity of effort or persistence students put into a task (Barrett, 2006; Russell, & Barrett, 1999). *Cognitive engagement* is defined as the quality of students' thinking while they are engaged in the tasks.

Reviewing lab studies in psychology and across several of their own projects, Linnenbrink (2007) examined the extent to which pleasant and unpleasant emotions have contributed to different forms of engagement with somewhat inconsistent findings. She reviewed several laboratory studies of unpleasant emotions where researchers found that unpleasant affect led students to attend to the task/details more carefully. However, Linnenbrink, Ryan, and Pintrich (1999; see also Linnenbrink & Pintrich, 2002) found in their own work that unpleasant emotions were negatively related to students' working memory functioning and learning. Findings



from this model led us to wonder if apparently inconsistent findings regarding the connection between emotion and learning may reflect the interplay of activating emotions and the strategies they activate. In other words, for some students, activating unpleasant emotions during tests may lead to enhanced performance because those emotions activate students to engage strategies that are task focused. On the other hand, for some students, activating unpleasant emotions during tests may impede performance because those emotions may activate students to use less productive, more emotion-focused strategies. Thus, according to Linnenbrink's research, students who evidence anxiety during tests should also evidence different patterns of coping compared with students who do not (Barrett, 2006; Russell, & Barrett, 1999). Moreover, inconsistent findings across the literature on test performance suggest that there may be multiple patterns of coping and performance, with higher performing groups activating adaptive task-focusing and regaining-task-focusing processes and others activating more emotion-focused processes.

### *Control-Value Theory of Achievement Emotions*

In their control-value theory of achievement emotions, Pekrun et al. (2007) defined *achievement emotions* as those "tied directly to achievement activities or achievement outcomes" (p. 15). Pekrun and colleagues (Pekrun, 2000; Pekrun et al., 2002, 2007) developed a taxonomy of achievement emotions along three dimensions: valence, activation, and control. Aligned with Linnenbrink (2007), their framework posits that anxiety is an outcome-oriented achievement emotion that has an unpleasant valence and serves an activating function. However, they further argued that any exploration of achievement emotions must include the dimension of control. Thus, control-value theory adds to the understanding of systematic differences in students' approaches toward tests by positing that students who perceive tests as within their control are likely to evidence different patterns of anxiety, coping, and achievement.

Pekrun and colleagues (Pekrun, 2000; Pekrun et al., 2002, 2007) outlined three types of subjective control expectancies: action-control expectancies, action-outcome expectancies, and situation-outcome expectancies. Action-control expectancies include students' belief that "an achievement activity can be successfully initiated and performed" (Pekrun et al., 2007, p. 18). These types of expectancies parallel the self-efficacy types of beliefs that Bandura (1977) described in his theory of personal agency. Action-outcome expectancies include students' beliefs that their "activities lead to outcomes one wants to attain" (Pekrun et al., 2007, p. 18). From the perspective of cognitive appraisal theory, action-outcome expectancies parallel agency/control appraisals (Schutz & Davis, 2000; Smith, 1991). Finally, situation-outcome expectancies reflect beliefs "that these outcomes occur in a given situation without one's action" (Pekrun et al., 2007, p. 18). Thus, from the perspective of cognitive appraisal theory, action-outcome expectancies parallel future-expectancy appraisals (Smith, 1991). In their research, Pekrun et al. (2007) identified that dimensions of perceived control and perceived value were both necessary to predict which achievement emotions were likely to arise. Students' perceptions of control may also have important implications for the types of strategies they select to regulate their emotions (Boekaerts, 2007). Specifically, students who perceive themselves as

being in control of their test scores may be more likely to select more task/problem-focused coping strategies. Likewise, do students who perceive themselves as not in control of their test scores select more emotion-focused coping strategies?

Consistent with work by Gross (2002) and Lazarus (1999), Pekrun et al. (2007) outlined four targets of students' efforts to regulate emotions. They argued that students can gear their regulation efforts toward managing the emotion itself, toward the appraisals that underlie the emotions and the competencies that underlie students' judgments of agency, or toward modifying the environment. What is unique about their model of regulation is their discussion of how students might regulate their emotions by improving their competency. Consistent with Gross and Munoz (1995; Gross & Thompson, 2007), this behavior reflects a type of situation modification in which students attempt to assert themselves into the situation by modifying what they know or the strategies they use. Few, however, have studied the extent to which strategies used during the test may contribute to later competency building.

### *Dual-Processing Self-Regulation Model*

Boekaerts (1993; see also Boekaerts, 2007) outlined a dual-processing approach for understanding how students' appraisals of academic situations might lead to two different pathways: a growth pathway and a well-being pathway. In her theory, Boekaerts moved beyond a simple analysis of value and importance to argue that the critical appraisal students make about their academic tasks is the extent to which the task is congruent with their goals (i.e., whether the task aligns with their personal and academic goals). Academic tasks may be important for a variety of reasons, but the extent to which tasks align with students' goals can set them on two entirely different self-regulatory pathways. In the growth pathway, students appraise the academic task as aligned with their goals, and they are more likely, when faced with stress, to select strategies that refocus them on the learning goals. In contrast, in the well-being pathway, students appraise the mismatch (i.e., incongruence) between the task and their goals as a warning signal that something is "wrong" (Boekaerts, 2007, p. 39). When faced with stress, students are more likely to select strategies (e.g., avoiding, distancing, working to preserve one's sense of self) that redirect their attention away from learning goals. Boekaerts (1993, 2007) noted there are likely to be individual differences in the amount of unpleasantness, or stress, students are likely to tolerate in academic tasks—with some students being more prone to switch paths from growth to well-being, because they cannot tolerate much stress. Dual-processing theory contributes to the understanding of how students' judgments of congruence can lead them to select coping strategies that are more or less oriented toward growth.

From this perspective, we expected to find groups of students evidencing a growth pathway, whereby appraisals of tests as goal congruent would be associated with selecting coping strategies oriented toward growth, that is, more task-focused strategies. Although habitual appraisals of tests as goal incongruent would be associated with selecting coping strategies that preserve well-being, this is wishful thinking and importance reappraisal. Moreover, we wondered if students who were less tolerant of stress in



testing situations could be identified on the basis of their patterns of appraising tests.

### Purpose

The overall purpose of this study was to explore whether groups of students share similar patterns of appraising tests and to examine whether these patterns evidenced different consequences for students' (a) experience of anxiety during test taking, (b) deployment of emotion-regulation strategies, and (c) academic achievement. Ultimately, we hope to examine the ways in which some students' appraisals of tests, in general, may make them more prone to experience state anxiety during tests. Furthermore, we hypothesize that students' deployment of strategies to regulate anxiety during tests may be, in part, a reflection of their beliefs about tests and about themselves as test takers. This is an exploratory study in that (a) to our knowledge, there has never been a clustering of students on appraisal dimensions nor has there been an examination of the coping strategies associated with different patterns and (b) we had no *a priori* hypotheses as to how many clusters we would find and how patterns of judgments would cluster together. On the other hand, there is a growing body of literature from which to make hypotheses about what types of differences we might find. This is where we drew from the circumplex (Barrett, 2006; Linnenbrink, 2007; Russell, & Barrett, 1999), control value (Pekrun et al., 2007), and dual-processing (Boekaerts, 2007) theories to make general suggestions about differences.

We drew from Schutz and colleagues' (Schutz & Davis, 2000; Schutz & DeCuir, 2002; Schutz et al., 2004, in press) theory and measures of emotion regulation during testing because they contained dimensions/subscales that captured the different judgments implicated by the literature. In many ways, their theory represents a bridge across the other three theories in that it contains two dimensions of valence (importance and congruence) and, in doing so, may enable us to explore the extent which students' appraisals of test taking lead them to have qualitatively different types of engagement during tasks. Additionally, their scale includes two different dimensions of control expectancies: action control (i.e., the agency scale) and action outcome (i.e., the testing problem efficacy scale). By examining two dimensions of control, we may come to better understand why students choose to deploy some strategies over others, as well as the functions each strategy may serve in regulating emotions during testing.

This project consisted of three phases of analysis. In Phase 1, our goal was to identify a theoretically defensible cluster solution that would replicate in a second independently clustered sample. Specifically, clustering students along the four appraisal dimensions, we identified a five-cluster solution that appeared both empirically and theoretically defensible. In Phase 2, our goal was to see if the solution identified in Phase 1 would replicate in a second sample. Specifically, we clustered Sample 2 into five clusters and compared across the solutions to see if the same clusters appeared. In the event of parallel solutions, we then submitted both samples to a cross-classification validation analysis. This type of analysis generates a prediction rule off the first sample and applies it to classify the second sample. Classification assignments can then be compared with the independent clustering assignment generated for Sample 2. In Phase 3, our goal was to examine the extent to which different patterns of

appraising tests predicted differences in anxiety and emotion-regulation strategies. Specifically, in this analysis, we used a five-cluster solution built from the entire data set ( $N = 2,215$ ), generated theoretically driven hypotheses about the differences we might find, and completed a series of analyses to externally validate the five-cluster solution.

### Method

#### Participants

Prior to the start of their first year at college, the incoming class ( $n = 5,872$ ) of first-year students (ages 18–19 years) at a large Southwestern university completed a battery of assessments, including the Learning and Study Strategies Inventory (Weinstein & Palmer, 2002; Weinstein, Palmer, & Hanson, 1995; Weinstein, Palmer, & Schulte, 1987), which is a measure of learning strategies used and test anxiety experienced while in high school, and the Emotional Regulation During Test Taking Scale (ERT; Schutz et al., 2004).<sup>2</sup> The sample included slightly more female (56%) than male (44%) students. In addition, data from the state department of education were compiled for all students, including Verbal and Quantitative scores for the Scholastic Aptitude Test (SAT), percentile rank in high school, and first-semester college grade point average. Students with incomplete profiles were excluded from analyses. The final sample size for this study was 2,215.

#### Measures

To identify homogeneous groups of students on the basis of their patterns of appraisal, we submitted undergraduate students' responses to Cognitive-Appraising Processes subscales of the ERT (Schutz et al., 2004) to cluster analysis. Students were asked to think about tests and test taking in general when responding to ERT items. Responses were given on a Likert-type response scale with five options: 1 (*almost never*), 3 (*sometimes*), and 5 (*almost always*); Options 2 and 4 were unlabeled.

*Capturing students' cognitive appraisals for tests.* Items on the Cognitive-Appraising Processes subscales of the ERT are designed to measure the four discrete judgments students may make about tests. These subscales include Goal Importance (4 items), Goal Congruence (4 items), Agency (4 items), and Testing Problem Efficacy (4 items). Goal Importance items ask students to identify the extent to which test taking is important for meeting their goals (e.g., "My test performances are important for getting my degree."). Our sample yielded an alpha of .79. Items on the Goal Congruence subscale are designed to capture the extent to which students perceive their grades on tests as helping them to accomplish their goals (e.g., "My grades on tests are helping me get my degree."). Schutz et al. (2004) reported internal consistency estimates of  $r = .80$  for Goal Congruence items. Our sample yielded an alpha of .76. Items on the Agency subscale capture the extent to which students perceive that they are in control of the outcome of their tests (e.g., "It is my own fault if I don't do well on tests."). Schutz et al. reported internal consistency estimates of

<sup>2</sup> Only the Test Anxiety subscale of the Learning and Study Strategies Inventory was used in the project. The other subscales fell outside the scope of this article.



$r = .85$  for agency items. Our sample yielded an alpha of .73. Finally, items on the Testing Problem Efficacy subscale evaluate the extent to which students feel confident they are capable of managing the problems that emerge during test taking (e.g., "Even when there are difficult questions, I can usually figure out how to answer them."). Schutz et al. reported internal consistency estimates of  $r = .79$  for the Testing Problem Efficacy subscale. Our sample yielded an alpha of .75.

*Concurrent measures.* In Phase 3, we examined the validity of our replicated cluster solution by examining the extent to which students with similar patterns of appraisal for tests experienced distinct patterns of anxiety, emotion regulation, and achievement. For these analyses, we examined differences among clusters, using information that was not included in the cluster analysis. These variables included students' scores on the Test Anxiety subscale of the Learning and Study Strategies Inventory, their scores on the coping dimensions of the ERT, and achievement data from their college records. The Test Anxiety scale of the Learning and Study Strategies Inventory (Weinstein et al., 1987) comprises nine items assessing the degree to which students worry about their academic performance. For example, items address the extent to which worrying about performance during a test tends to interfere with students' concentration.

Students' scores on this scale measure how tense or concerned they are when approaching academic tasks. Students who score low on this measure (indicating high anxiety) need to learn techniques for coping with anxiety and reducing worry so that they can focus on the task at hand and not on their anxiety. (Weinstein & Palmer, 2002, p. 9)

Weinstein et al. (1987) reported an internal consistency estimate of  $r = .87$ . Our sample yielded an alpha of .85.

The ERT comprises five types of processes organized around three dimensions of emotion regulation related to test taking. Four items were used to assess the task-focusing processes dimension (e.g., "I work harder to find the main idea in the questions."). Schutz et al. (2004) reported internal consistency estimates of  $r = .57$  for the Task Focusing subscale. Our sample yielded an alpha of .59. Two subscales attempted to capture the regaining-task-focusing processes dimension. These included Tension Reduction (4 items; e.g., "I try to slow down my breathing.") and Importance Reappraisal (5 items; e.g., "I try to keep the test's importance in perspective with the other factors in my life."). Schutz et al. reported internal consistency estimates of  $r = .77$  and  $r = .72$  for the Tension Reduction and Importance Reappraisal subscales, respectively. Our sample yielded alphas of .63 and .67, respectively. Last, two subscales were designed to capture an emotion-focused processes dimension. These included the Wishful Thinking subscale (4 items; e.g., "I hope the teacher decides not to count the test.") and the Self-Blame subscale (4 items; e.g., "I blame myself for the problems I am having on the test."). Schutz et al. reported internal consistency estimates of  $r = .77$  and  $r = .86$  for wishful thinking and self-blame subscales, respectively. Our sample yielded alphas of .77 and .83, respectively.

In addition, at the end of the survey, participants were asked to self-report their age and their gender. These data were merged with data from the institution's database containing students' SAT Verbal and Quantitative scores, their relative rank in high school, and their first semester college grade point average.

## Procedures

*Phase 1: Independent cluster analysis.* Following procedures outlined by DiStefano and Kamphaus (2006), the larger sample of students was divided into two randomly drawn subsamples: an initial sample (used for an initial cluster analysis;  $n = 1,107$ ) and a second sample ( $n = 1,108$ ) used for a second, independent cluster analysis and a cross-classification internal validation. Cluster analysis refers to a set of procedures used to uncover homogeneous groups underlying a data set (Aldenderfer & Blashfield, 1984; Anderberg, 1973; Blashfield & Aldenderfer, 1988; DiStefano & Kamphaus, 2006; Milligan & Cooper, 1987). Our goal was to identify smaller subgroups of first-year college students who are similar, with regard to their judgments of tests, to members within their cluster while distinct from members of the other clusters. To achieve this goal, we used a two-step procedure in our initial cluster analysis that combined Ward method and K-means algorithms to attempt to overcome the limitations of each method when selected as the sole method (DiStefano & Kamphaus, 2006). To guide our selection of a clustering solution for Sample 1, we began by examining the cubic clustering criterion (CCC) plots to determine the number of potential clusters underlying our data. This criterion indicated there might be between four and six clusters embedded in the data. Therefore, on the basis of these plots, we ran four- through seven-cluster solutions for the initial solution. Each solution was scrutinized to the extent that we could identify unique patterns of the four appraisals for tests. To facilitate interpretation, we graphed both raw and standardized scores. Although raw scores assisted us in understanding the magnitude of students' reports; standardizing the scores assisted in interpreting the relative magnitude of each cluster's reports. Once we failed to recognize unique solutions—in other words, one cluster appeared to split into two—we chose to stop interpreting new solutions. From these clustering procedures, we identified a five-cluster solution as optimal (see Table 1 and Figure 1).

*Phase 2: Replication and cross-classification.* To explore the validity of the five-cluster solution generated in Sample 1 ( $n = 1,107$ ), Sample 2 ( $n = 1,108$ ) was independently clustered into five clusters to determine whether the same five-cluster solution could be replicated. We repeated the clustering procedures from Phase 1, considering Sample 2 an independent analysis, and examined (a) whether the same number of clusters were present in the data and (b) whether the patterns of cognitive appraisal found in Sample 2 reflected patterns similar to those found in Sample 1 (see Table 2, Figure 2).

Next, a classification rule, developed from the initial five-cluster solution, was used to classify cases in Sample 2. Hit rates for correctly and incorrectly classified cases across the five clusters were compared (Huberty, 1994). High hit rates indicate that the rule generated from Sample 1 predicted students' actual cluster assignment for Sample 2. Findings from the cross-classification enabled us to examine the extent to which the rule developed in Sample 1 could replicate patterns found in Sample 2. The use of these two methods allowed us to examine the data for sample-specific characteristics to determine the extent to which clusters represented unique, stable groups that may exist in the larger population of first-year college students (See Table 3).

*Phase 3: Examining the corollaries of different patterns of appraising tests.* To explore the external validity of the five-cluster solution, we examined whether uniquely different patterns

Table 1  
Cluster Centroid Values for the Initial Five-Cluster Appraisal Solution

Test appraisal	Cluster centroid value					Sample mean
	Cluster 1 (n = 199)	Cluster 2 (n = 257)	Cluster 3 (n = 277)	Cluster 4 (n = 155)	Cluster 5 (n = 219)	
Goal relevance	15.32 (0.68)	11.77 (-0.43)	12.51 (-0.20)	8.54 (-1.44)	16.87 (1.16)	13.15
Goal incongruence	15.20 (0.50)	12.81 (-0.30)	12.23 (-0.16)	9.37 (-1.46)	17.08 (1.13)	13.71
Agency	16.68 (0.19)	17.75 (0.62)	13.82 (-0.95)	14.97 (-0.49)	17.81 (0.65)	16.20
Testing problem efficacy	10.88 (-0.98)	15.46 (0.74)	12.67 (-0.31)	12.18 (-0.50)	15.57 (0.77)	13.50

Note. Standardized scores are presented beside each raw value in parentheses.

of cognitive appraisals during test taking were associated with uniquely different patterns of anxiety, emotion regulation, and achievement by examining variables, which were not included in the clustering process. To determine the extent to which groups were different, one-way analyses of variance (ANOVAs) were conducted between clusters where anxiety, coping, and achievement measures (i.e., SAT, high school and college standing) were considered the dependent variables. If an omnibus ANOVA test illustrated statistically significant differences among the clusters, follow-up tests were run with Tukey's post hoc procedures, controlling the error rate to .05, to identify statistically different (as well as similar) clusters. For variables in which the assumption of homogeneity of variance was violated, Dunnett's post hoc test for unequal variances was used.

## Results and Discussion

### Phase 1: Initial Cluster Analysis

Findings from the independent cluster analysis are presented in Table 1. To assist with comparing what makes each cluster unique, Figure 1 portrays each cluster's standardized score along the four appraisals. Regarding first-year college students' cognitive appraisals about tests, our findings yielded five unique patterns of appraisals. Here we describe each cluster.

*Cluster 1.* In general, students in this cluster were characterized by moderate levels of goal importance, goal congruence, and agency. According to their raw data, they perceived tests as important and their scores on tests as generally helping them to achieve their goals. Relative to the other clusters, their appraisals

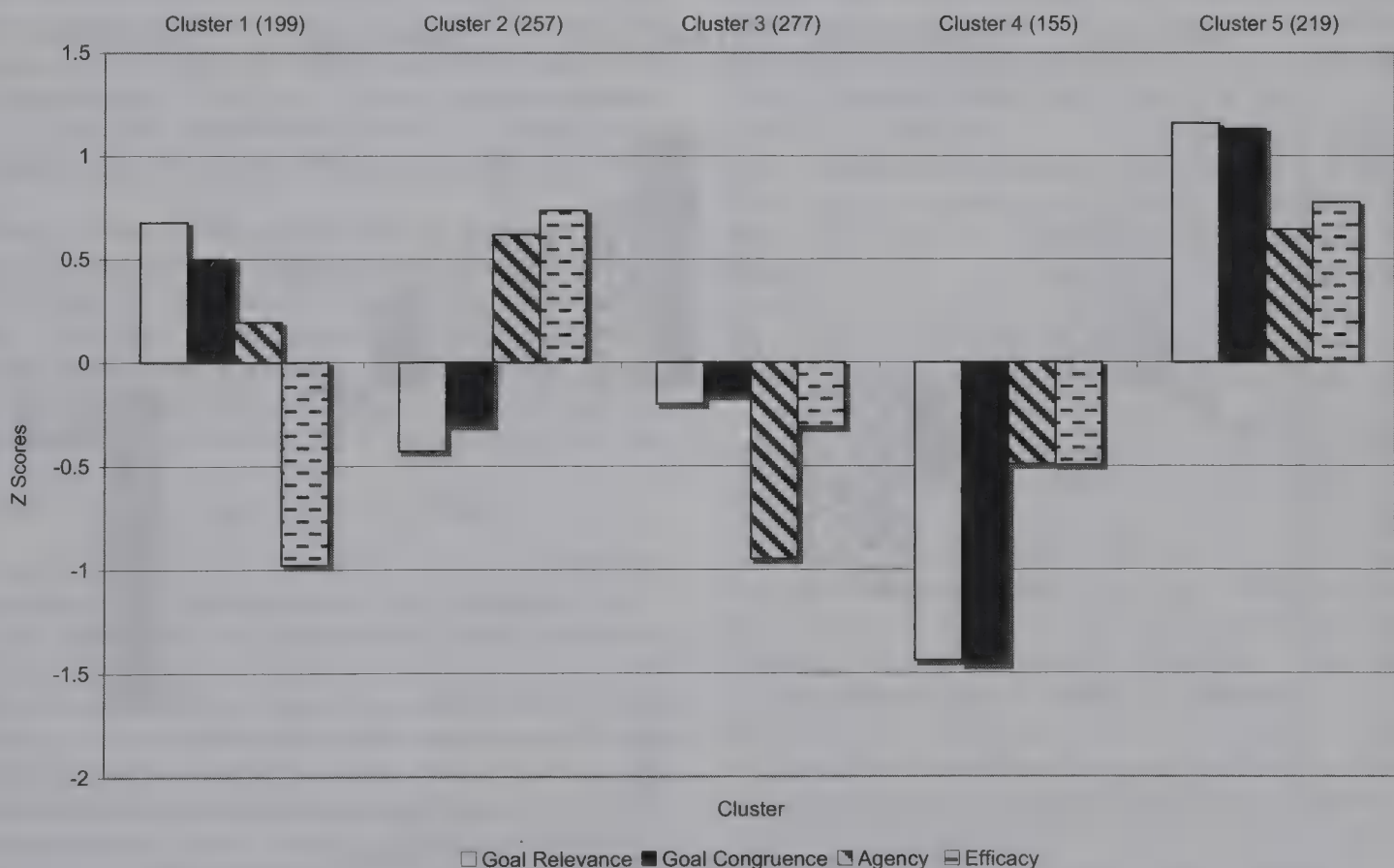


Figure 1. An examination of patterns of cognitive appraisals for the initial five-cluster solution (Z scores). To assist with comparison, scores on the four cognitive appraisal scales were converted to standardized scores. Note that higher scores on the Learning and Study Strategies Inventory Test Anxiety scale reflect lower rates of test anxiety.



Table 2  
Cluster Centroid Values for the Replication Five-Cluster Appraisal Solution

Test appraisal	Cluster centroid value					Sample mean
	Cluster 1 (n = 264)	Cluster 2 (n = 268)	Cluster 3 (n = 214)	Cluster 4 (n = 166)	Cluster 5 (n = 196)	
Goal relevance	15.17 (0.65)	12.39 (-0.20)	11.46 (-0.47)	8.17 (-1.47)	16.88 (1.16)	13.04
Goal incongruence	14.96 (0.44)	13.49 (-0.04)	12.26 (-0.44)	9.05 (-1.50)	17.31 (1.21)	13.61
Agency	15.90 (-0.15)	17.37 (0.42)	13.85 (-0.95)	15.63 (-0.26)	18.54 (0.87)	16.29
Testing problem efficacy	12.20 (-0.50)	15.69 (0.79)	11.37 (-0.81)	12.91 (-0.24)	15.40 (0.68)	13.56

Note. Standardized scores are presented beside each raw value in parentheses.

of goal importance and congruence were relatively high. However, their slightly higher levels of goal importance and congruence were paired with lower levels of agency and efficacy. Although these students tended to report feeling in control of their outcomes on tests, they reported feeling the least confident to manage problems that emerged during tests.

**Cluster 2.** In general, students in this cluster were characterized by moderate levels of all the appraisals. However, relative to the other clusters, their appraisals of goal importance and goal congruence were somewhat low and were paired with higher relative reports of agency and test problem efficacy. In fact, these students reported the highest levels of both agency and problem efficacy. Thus, although students in this cluster reported that their test scores did not always help them achieve their goals, they

perceived that they were in control of their outcome and were able to manage the problems that emerged.

**Cluster 3.** Although students in Cluster 3 reported relatively moderate levels of goal importance, goal congruence, and problem efficacy, they were defined by having the lowest levels of agency. Thus, students in this cluster believed tests were relatively important, that their scores generally helped them accomplish their goals, and that when faced with problems on tests they were somewhat able to manage them. However, students in this cluster reported that they generally did not feel in control of their test scores.

**Cluster 4.** These students were characterized by the lowest rates of goal importance and congruence. Thus, students in this cluster generally did not view tests as important to their goals, nor

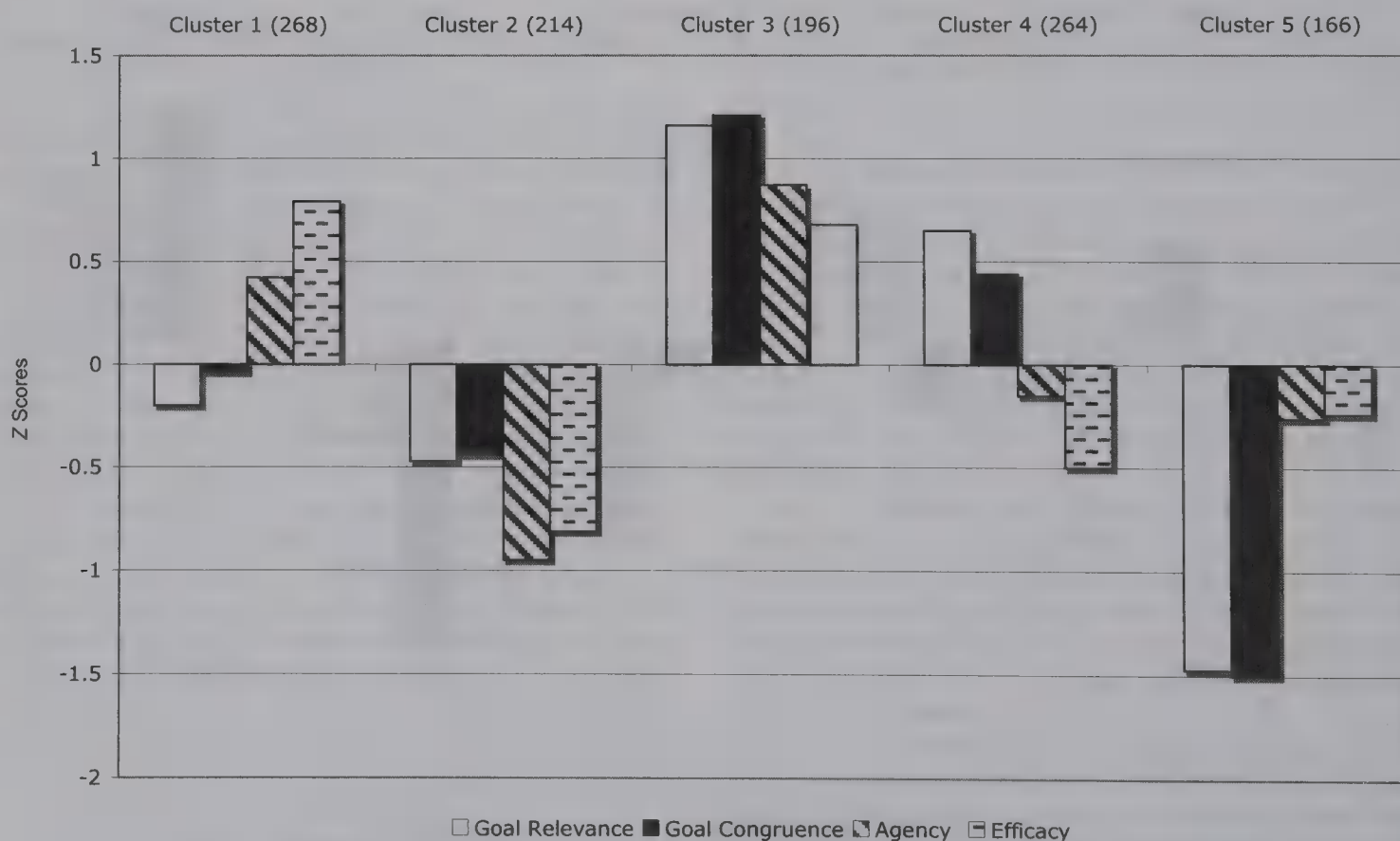


Figure 2. An examination of patterns of cognitive appraisals for the five-cluster replicated solution (Z scores). To assist with comparison, scores on the four cognitive appraisal scales were converted to standardized scores. Note that higher scores on the Learning and Study Strategies Inventory Test Anxiety scale reflect lower rates of test anxiety.

Table 3  
*Examining Cross-Classification Findings for Replication Sample 2 With a Classification Rule  
 Generated From Initial Sample 1*

Validation sample: Cluster classification by rule	Validation sample: Independent clustering assignment					Total
	1	2	3	4	5	
Cluster 1						
Raw count	<b>152</b>	0	9	0	10	171
%	<b>57.6</b>	0.0	4.2	0.0	5.1	15.4
Cluster 2						
Raw count	2	<b>226</b>	3	36	0	267
%	0.8	<b>84.3</b>	1.4	21.7	0.0	24.1
Cluster 3						
Raw count	82	33	<b>171</b>	1	0	287
%	31.1	12.3	<b>79.9</b>	0.6	0.0	25.9
Cluster 4						
Raw count	0	0	31	<b>129</b>	0	160
%	0.0	0.0	14.5	<b>77.7</b>	0.0	14.4
Cluster 5						
Raw count	28	9	0	0	<b>186</b>	223
%	10.6	3.4	0.0	0.0	<b>94.9</b>	20.1
Total						
Count	264	268	214	166	196	<b>1,108</b>
%	100.0	100.0	100.0	100.0	100.0	100.0

*Note.* Columns = assignments of validation sample (Sample 2 data) by independent clustering method; rows = assignment of validation sample (Sample 2) based on the classification rule developed from Sample 1. % = percentages within validation sample for independent clustering. Values in bold indicate the hit rate for each cluster, that is, the number and relative percentage of students who were accurately classified with the prediction rule.

did they view their scores as helping them accomplish their goals. It is interesting that, relative to the other clusters, the students in this cluster reported moderate levels of agency and congruence. Thus, although they did not tend to perceive tests as important or goal congruent, they reported feeling relatively in control of their outcomes and able to manage the challenges that arose during tests.

*Cluster 5.* Students in this cluster reported the highest levels of all four appraisals. Although their reports of problem efficacy were relatively high compared with the other clusters, relative to their reports of goal importance, congruence, and agency, their feelings of problem efficacy were surprisingly low. Thus, despite feeling in control of test outcomes, these students reported some uncertainty about their ability to manage problems that emerged during tests.

### *Phase 2: Replication and Cross-Classification*

To validate the solution, an independent sample (Sample 2) was first clustered to determine whether the same five clusters could be replicated. Again, the same procedures were used to cluster the data, with four- through seven-cluster solutions run and interpreted. On the basis of the independent sample analysis, we felt confident that a five-cluster solution was again optimal. Findings from the independent clustering of Sample 2 are presented in Table 2. To assist the reader with interpretation, clusters in the replication sample (Sample 2) were named to be aligned and presented in the same order as the solution generated in the initial cluster (Sample 1).

Regarding the interpretation of the new five-cluster solution (see Table 2), we felt confident that the substantive patterns of appraisals and emotion regulation existed in both samples. Four of the five

clusters were easily identifiable and evidenced nearly identical patterns of appraisal across the two samples. The only cluster we had some concern about was Cluster 1. On the one hand, this cluster evidenced a similar relative pattern of cognitive appraisals, with generally moderate levels of goal importance, goal congruence, and agency and relatively higher reports of goal importance and congruence compared with agency and problem efficacy. What did not appear to replicate from the first sample were the lower reports of test problem efficacy. Thus, in the replication sample, students did tend to report being among the least confident for dealing with the problems that emerge during tests. However, their reports, relative to the initial sample, were not as low.

Next, to provide additional validation evidence, we used a cross-classification procedure to determine if a classification rule developed from the initial five-cluster solution could accurately classify students into the five groups as well as independent clustering. The extent to which both solutions (i.e., cross classification and independent clustering) produce similar results supports the existence of the five clusters of appraisal. As previously discussed, it is common to apply a classification rule built on an existing cluster solution to classify ungrouped cases. The classification rule is based on predictive discriminant analysis in which the linear classification function is used to predict membership of an ungrouped case by assigning each case to the cluster with which it is most closely associated (Huberty, 1994; Huberty, DiStefano, & Kamphaus, 1997). Prediction of cluster membership requires information from the original sample, including cluster membership of the grouped cases and probability information relaying the likelihood of encountering a certain subgroup of children in the population. The classification rule uses a combination of the def-



inition of the cluster solution and prior probability information to classify a case into the cluster with which it is most closely associated. We may compare results between cases assigned with the classification rule and those assigned through independent clustering to determine if children were assigned to similar clusters. If cross-classification results for students classified with the rule built on the initial sample are similar to results from independent clustering, this provides confidence that a student would have received the same or a similar cluster assignment regardless of the grouping method used to classify the data. The use of these two methods allowed us to examine the data for sample-specific characteristics to determine the extent to which clusters represent unique, stable groups that may exist in the larger population.

To examine the extent to which the rule developed in Sample 1 could replicate patterns found in Sample 2, a  $5 \times 5$  table was constructed and is reported in Table 3. In Table 3, rows correspond to the cluster solution when cases were assigned with the classification rule. Columns in Table 3 correspond to the cluster assignments achieved with the independent sample solution. The table can be interpreted by reading down a column to determine categorization of students by use of the classification rule built on the initial sample, or it can be interpreted by reading across a row to determine categorization by independent clustering of Sample 2. The hits along the main diagonal can be calculated and compared with the misses or off-diagonal elements. Hit rates were computed by taking the number of cases classified into the same cluster by both methods (i.e., those correctly classified), divided by the cluster sample size from the independent clustering. Again, high hit rates indicate the rule generated from Sample 1 predicted students' membership accurately with their actual cluster assignment for Sample 2.

To assist in judging the adequacy of results, criteria were developed by Huberty (1994; see also Huberty et al., 1997) to guide evaluation. Hit rates showing at least 75% agreement between the two classification methods are thought to represent high agreement between the two methods, hit rates with between 50% and 74% agreement represent moderate agreement, hit rates with between 30% and 49% agreement denote fair agreement, and hit rates with lower than 30% agreement represent poor agreement between the classification methods.

The cross-classification analysis detailed in Table 3 showed that most of the students were correctly classified into their assigned group. In fact, four of our five clusters evidenced a high degree of agreement across the two samples. Only one cluster, Initial Cluster 1, evidenced moderate levels of agreement. In part, these lower rates may reflect that the rule was generated on a cluster solution that evidenced much lower reports of testing problem efficacy than in the replication sample. Thus, students who did not fit this rule were most likely to be misclassified into Cluster 3. This is because the students in Cluster 3 similarly reported lower levels of testing problem efficacy.

### *Phase 3: Examining the Corollaries of Different Patterns of Appraising Tests*

Again, the purpose of these analyses was twofold. First, because we were able to find the same five-cluster solution in both samples, we wanted to verify that when the two samples were pooled, the same solution would emerge. In some instances, when samples are

pooled together, new clusters may emerge (DiStefano, Kamphaus, Horne, & Windsor, 2003). Thus, finding the same five clusters again would indicate the robustness of the solution. In addition, by pooling the samples together, we might reduce the sample-specific variations that led to the lower hit rates for Cluster 1.

Second, given our ability to replicate the solution, we wanted to examine whether the five uniquely different patterns of cognitive appraisals during test taking were associated with uniquely different patterns of anxiety, emotion regulation, and achievement. To determine the extent of differences among groups, one-way ANOVAs were conducted between clusters where anxiety, coping, and achievement measures (i.e., SAT, high school and college standing) were considered as the dependent variables. To minimize the sample-specific characteristics, particularly for the one cluster that evidenced slightly lower hit rates, we pooled the two samples together, using the assignments/labels from the initial sample as the referent group. Prior to running these analyses, we made a priori hypotheses about the differences we might find.

We began this phase of the analysis by pooling the two samples together and conducting a third, independent cluster analysis. Again, we identified a five-cluster solution as the best fit and determined that the same five clusters from Phase 1 and Phase 2 existed in the pooled sample (see Table 4, Figure 3). With our increased confidence about the existence of the five-cluster solution, we proceeded with naming and describing each cluster. Then, we moved forward with making some a priori hypotheses about distinctions we might find in the ways these students approached regulating unpleasant emotions, like anxiety, that emerge during tests.

### *Hypothesizing Differences in Patterns of Anxiety, Emotion Regulation, and Achievement*

According to Schutz and colleagues' (Schutz & Davis, 2000; Schutz et al., 2004, in press) theory of emotion regulation during test taking, students who perceive tests as relevant to accomplishing their goals, their scores as obstacles to achieving their goals (i.e., incongruent), and themselves as out of control with regard to their scores (i.e., agency) and managing problems during the test (i.e., problem efficacy) are likely to experience anxiety during tests. From this perspective, we hypothesized that two clusters—Clusters 1 and 3—were likely to evidence higher rates of anxiety, relative to the other clusters. This is because both clusters reported that they perceived tests as important and scores as somewhat congruent but also perceived loss of agency or lower efficacy in managing problems during tests. Additionally, we were curious about how much anxiety students in Cluster 5 reported. Although these students reported the highest rates of goal congruence, they also reported corresponding lower perceptions of agency and efficacy. We wondered if this cluster might include students with much higher standards for performance who may experience anxiety when faced with having to think, or reason through, challenging questions on tests.

Drawing from circumplex theory (Barrett, 2006; Linnenbrink, 2007; Russell & Barrett, 1999), Pekrun et al.'s (2007) control-value theory, and Boekaert's (2007) dual-processing theory, we hypothesized specific differences we might find across the clusters in anxiety, regulation, and achievement. Linnenbrink (2007) argued that anxiety is a negatively valenced, activating emotion. From this perspective, we hypothesized that Clusters 1 and 3

Table 4  
Correlations Among Study Variables

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Goal importance	—													
2. Goal congruence	.727**	—												
3. Agency	.264**	.307**	—											
4. Test problem efficacy	.126**	.192**	.311**	—										
5. Task focusing	.158**	.216**	.250**	.349**	—									
6. Importance reappraisal	-.142**	-.198**	-.082**	.082**	.113**	—								
7. Tension reduction	.123**	.170**	.161**	.207**	.506**	.189**	—							
8. Wishful thinking	-.061**	-.073**	-.127**	-.331**	.068**	.150**	.029	—						
9. Self-blame	.095**	.066**	.096**	-.283**	.016	-.025	-.012	.452**	—					
10. Test anxiety <sup>a</sup>	-.004	.093**	.191**	.550**	.117**	-.045*	.067**	-.499**	-.484**	—				
11. SAT-Verbal	-.063**	-.039	.102**	.277**	.132**	-.069**	.045*	-.130**	-.093**	.271**	—			
12. SAT-Quantitative	.008	.030	.068**	.260**	.033	-.142**	-.054*	-.116**	-.091**	.214**	.489**	—		
13. High school percentile rank	.035	.050*	.046*	.065**	.076**	-.018	.029	-.019	-.010	.021	.107**	.138**	—	
14. Cumulative first-semester college GPA	-.017	.018	.004	.072**	.078**	-.033	.052*	-.042*	-.029	.075**	.303**	.334**	.355**	—

Note. SAT = Scholastic Aptitude Test; GPA = grade point average.

<sup>a</sup>According to the scoring of the Learning and Study Strategies Inventory, low scores on anxiety are indicative of high test anxiety.

\*  $p < .05$  (two tailed). \*\*  $p < .01$  (two tailed).

would report patterns of regulation during tests that evidence active coping. Active coping on the ERT scale might reflect the deployment of task-focusing and/or tension-reduction strategies. Pekrun et al. argued that students who perceive themselves as being in control of their test scores may be more likely to select more task/problem-focused coping strategies. Thus, we hypothesized that we would find higher rates of task-focusing processes in Clusters 2 and 5. Pekrun et al. also outlined how students might gear their regulation efforts at managing the appraisals that underlie the emotions and the competencies that underlie students' judgments of agency. From this perspective, students who perceive themselves to be out of control, with regard to their tests scores (i.e., action outcome) or problems during tests (i.e., action control), might target their regulation strategies toward reappraising the importance of the tests. Specifically, we hypothesized that we would find higher reported use of importance reappraisal in Clusters 3 and 4. Boekaerts argued that students who perceive tests as obstacles to achieving their goals would choose regulation strategies that attempt to preserve their sense of well-being, whereas students who perceived tests as helping them achieve their goals would be select strategies that create opportunities for further growth. On the basis of Boekaert's theory, we identified two clusters that appeared to be on a growth path: Cluster 2 and Cluster 5. We hypothesized that students in these clusters would report more use of strategies promoting task focusing and regaining task focusing. In contrast, we hypothesized that students in Clusters 1, 3, and 4 might select strategies oriented toward preserving their sense of well-being including, perhaps, some strategies for regaining task focusing, like tension reduction and emotion-focusing processes. Finally, we felt confident we would find higher rates of achievement (as evidenced by standardized scores, rank, and grade point average) in Clusters 2 and 5. This is because students in these clusters generally perceived tests as helping them to accomplish their goals, and they felt more or less in control of tests.

A correlation matrix of all of the variables included in Phase 3 is presented in Table 4. Findings from the ANOVAs are presented in Table 5 and Figure 4. Several general trends emerged. First, in general, clusters did appear to have distinct patterns of anxiety- and emotion-regulation strategies. In many cases, at least three different levels (i.e., high, moderate, low) emerged. This was the case for test anxiety, wishful thinking, and some of the academic achievement data. In some cases, statistically significant differences were found across four levels, for example, in the case of task-focused and importance-reappraisal strategies. Second, although we did find some differences in the achievement data, particularly in our ability to distinguish between the higher achieving groups (Clusters 2 and 5) and the lower achieving groups (Clusters 1 and 3), all of the students in our sample tended to be academically talented. In other words, at the end of their first semester in college, the average grade for each group was right around a B.

Several of our hypotheses about cluster differences were confirmed. In the following sections, we review our hypotheses as they relate to the data in Table 5 and Figure 4 and describe the ways in which students' deployment of strategies may be, in part, a reflection of their appraisals about tests.

*Cluster 1: Tests out of control.* We hypothesized that students in this cluster would report experiencing anxiety during tests and would evidence active coping but not coping oriented toward managing the problems of the test. In other words, we expected to



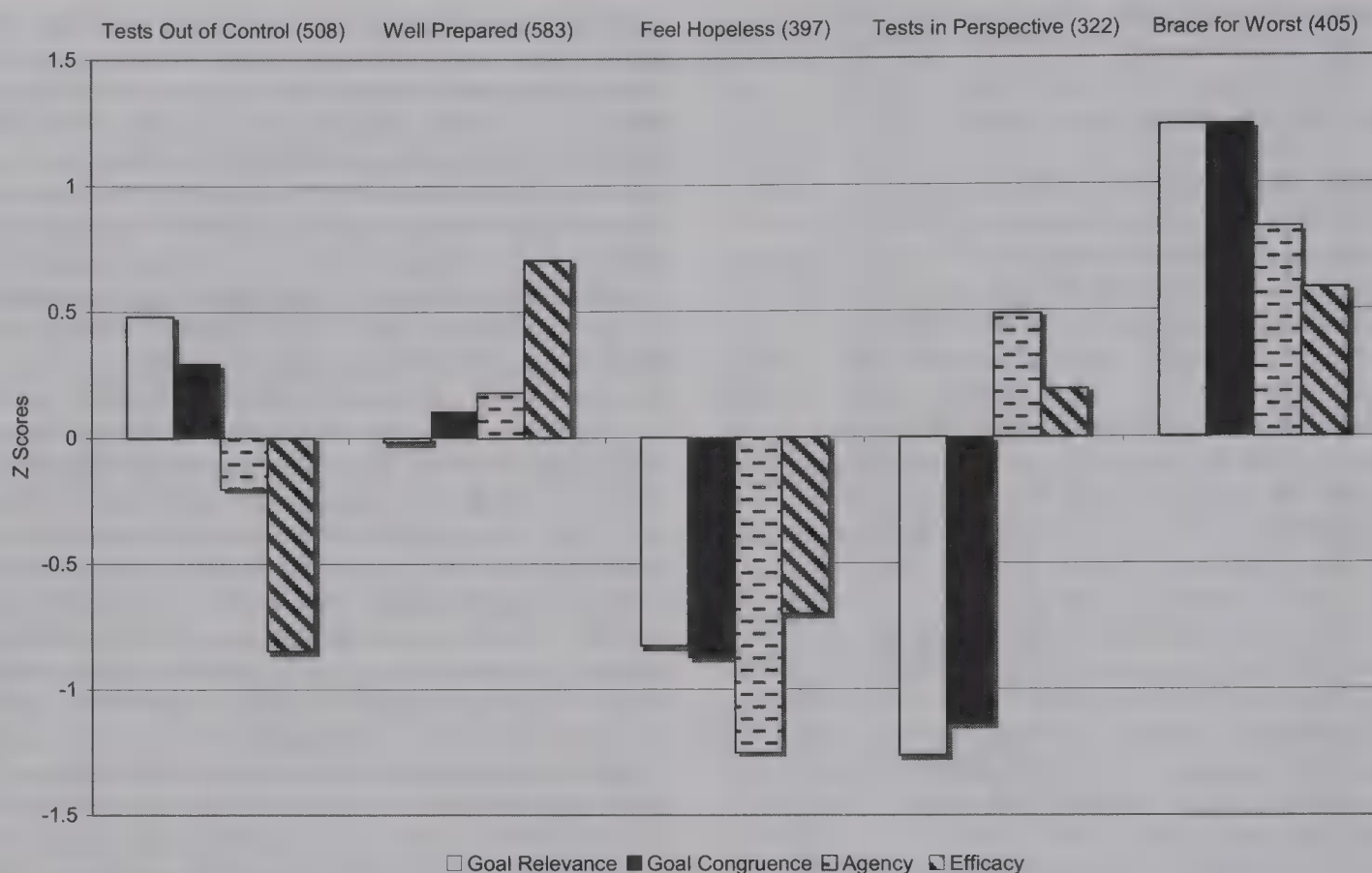


Figure 3. Final pooled sample five-cluster solution. Clusters were renamed when graphed to align with the clusters identified in Phase 1 of the analysis. To assist with comparison, scores on the four cognitive appraisal scales were converted to standardized scores. Note that higher scores on the Learning and Study Strategies Inventory Test Anxiety scale reflect lower rates of test anxiety.

find evidence suggesting that students in this cluster were on a well-being path. Specifically, we found that students in this cluster reported some of the highest anxiety during tests. When faced with problems during testing, students in this cluster reported lower rates of task-focusing and regaining-task-focusing regulation strategies paired with the highest rates of emotion-focusing strategies.<sup>3</sup> Thus, although they reported some of the highest levels of goal relevance and reported moderate levels of goal congruence, when faced with problems during tests, these students used the fewest strategies to manage the task and their emotions. These findings corroborate both the control-value and dual-processing theories of emotion. Finally, consistent with Linnenbrink et al. (1999; Linnenbrink & Pintrich, 2002), for students in this cluster, experiencing unpleasant emotions (i.e., anxiety) during tests was related to students' use of more emotion-focused strategies. In addition, when examining the achievement data, we noted that this group evidenced some of the lowest scores on standardized and college achievement measures.

**Cluster 2: Well prepared for challenges.** We hypothesized that students in this cluster would report lower levels of anxiety, would evidence patterns of regulation oriented toward growth (Boekaerts, 2007), and would have higher levels of achievement. Not surprisingly, students in Cluster 2 reported some of the lowest levels of test anxiety paired with reports of some of the highest standardized scores and college achievement. As hypothesized, the students in this cluster appeared to react coolly and calmly to test problems by deploying strategies oriented toward growth. This is consistent

with the control-value (Pekrun et al., 2007) theory of emotions, which states that students who perceive themselves as in control are more likely to select task-oriented strategies. When faced with problems during a test, students in this cluster reported some of the highest rates of task-focused strategy use paired with moderate rates of regaining-task-focusing strategies and lower rates of emotion-focused strategies.

**Cluster 3: Feeling hopeless.** As with Cluster 1, we hypothesized that students in this cluster would report experiencing anxiety during tests and would evidence active coping but not coping oriented toward managing the problems of the test. As hypothesized, students in this cluster evidenced some of the highest reports of anxiety during tests. Moreover, students in this cluster reported the lowest levels of task-focusing or tension-reduction strategies. Instead, students in this cluster appeared to rely on distancing strategies, reporting some of the highest levels of importance reappraisal and wishful thinking. In line with Boekaerts's (2007) dual-processing theory, we believe students in this cluster, when faced with challenges on tests, may have chosen to engage in self-preservation strategies by distancing themselves psychologically from the test. When looking at achievement data, it may be hard to think of students with a B average as having more chronic problems during tests; however, it is important to remember that

<sup>3</sup> Note that higher scores on the Learning and Study Strategies Inventory Test Anxiety scale reflect lower rates of test anxiety.

Table 5

*Examining Differences in Students' Test Anxiety, Emotion Regulation, and Achievement as a Function of Their Unique Patterns of Appraising Tests*

Emotion regulation, achievement, and gender	Tests out of control ( <i>n</i> = 508)	Well prepared ( <i>n</i> = 583)	Feeling hopeless ( <i>n</i> = 397)	Tests in perspective ( <i>n</i> = 322)	Bracing for the worst ( <i>n</i> = 405)
Test appraisal					
Goal relevance					
<i>M</i>	14.66 <sub>4</sub>	13.04 <sub>3</sub>	10.40 <sub>2</sub>	8.98 <sub>1</sub>	17.11 <sub>5</sub>
<i>SD</i>	1.72	1.56	2.07	2.22	1.75
Goal congruence					
<i>M</i>	14.56 <sub>4</sub>	13.99 <sub>3</sub>	11.02 <sub>2</sub>	10.22 <sub>1</sub>	17.39 <sub>5</sub>
<i>SD</i>	1.77	1.59	2.13	2.14	1.67
Agency					
<i>M</i>	15.73 <sub>2</sub>	16.70 <sub>3</sub>	13.08 <sub>1</sub>	17.48 <sub>4</sub>	18.36
<i>SD</i>	2.04	1.92	1.94	1.79	1.56
Test problem efficacy					
<i>M</i>	11.26 <sub>1</sub>	15.42 <sub>4</sub>	11.64 <sub>2</sub>	14.03 <sub>3</sub>	15.10 <sub>4</sub>
<i>SD</i>	2.13	1.57	2.13	2.44	2.26
Anxiety during tests <sup>a</sup>					
Anxiety during tests					
<i>M</i>	25.90 <sub>1</sub>	31.62 <sub>3</sub>	26.60 <sub>1</sub>	30.33 <sub>2</sub>	30.57 <sub>2,3</sub>
<i>SD</i>	5.93	5.50	5.93	6.12	6.59
Emotion-regulation strategies					
Task focused					
<i>M</i>	15.05 <sub>2</sub>	16.07 <sub>3</sub>	14.57 <sub>1</sub>	15.93 <sub>3</sub>	16.81 <sub>4</sub>
<i>SD</i>	2.25	2.16	2.27	2.24	2.17
Importance reappraisals					
<i>M</i>	14.26 <sub>1,2</sub>	14.74 <sub>2,3</sub>	15.12 <sub>3,4</sub>	15.65 <sub>4</sub>	14.00 <sub>1</sub>
<i>SD</i>	3.04	3.11	3.11	3.62	3.31
Tension reduction					
<i>M</i>	13.80 <sub>1,2</sub>	14.66 <sub>3</sub>	13.52 <sub>1</sub>	14.29 <sub>2,3</sub>	15.20 <sub>4</sub>
<i>SD</i>	2.52	2.54	3.11	2.79	2.90
Wishful thinking					
<i>M</i>	14.17 <sub>3</sub>	12.05 <sub>1</sub>	14.03 <sub>3</sub>	13.06 <sub>2</sub>	12.49 <sub>1,2</sub>
<i>SD</i>	3.36	3.44	3.36	3.79	3.86
Self-blame					
<i>M</i>	13.98 <sub>4</sub>	12.26 <sub>1</sub>	13.08 <sub>2,3</sub>	12.84 <sub>1,2</sub>	13.62 <sub>3,4</sub>
<i>SD</i>	3.41	3.45	3.29	3.99	4.02
Academic data					
SAT-Verbal					
<i>M</i>	572.80 <sub>1</sub>	620.69 <sub>3</sub>	582.38 <sub>1</sub>	616.77 <sub>3</sub>	599.82 <sub>2</sub>
<i>SD</i>	77.26	79.48	77.89	83.90	85.15
SAT-Quantitative					
<i>M</i>	602.24 <sub>1</sub>	643.97 <sub>5</sub>	609.00 <sub>1,2</sub>	619.53 <sub>2,3</sub>	630.28 <sub>3,4</sub>
<i>SD</i>	77.36	80.03	80.10	79.32	85.36
High school percentile rank					
<i>M</i>	87.14	88.21	86.50	86.75	88.28
<i>SD</i>	10.90	10.26	11.06	10.89	10.33
Cumulative first-semester GPA					
<i>M</i>	3.01 <sub>1</sub>	3.17 <sub>2</sub>	3.07 <sub>1,2</sub>	3.08 <sub>1,2</sub>	3.06 <sub>1,2</sub>
<i>SD</i>	0.67	0.70	0.63	0.68	0.72
Gender					
Male					
<i>N</i>	159	284	123	131	185
%	31.3	48.7	31.0	40.7	45.7
Female					
<i>N</i>	349	299	274	191	220
%	68.7	51.3	69.0	59.3	54.3

*Note.* Tukey's post hoc tests are denoted by numbered subscripts in the table. When interpreting the post hoc tests, it is noted that lower numbers (e.g., 1) report lower mean scores for a given variable than higher numbers (e.g., 4). Also, clusters with the same number were not statistically different; however, clusters with different numbers were statistically different (see Table 4). SAT = Scholastic Aptitude Test; GPA = grade point average.

<sup>a</sup> According to the scoring of the Learning and Study Strategies Inventory, low scores on anxiety are indicative of high test anxiety.



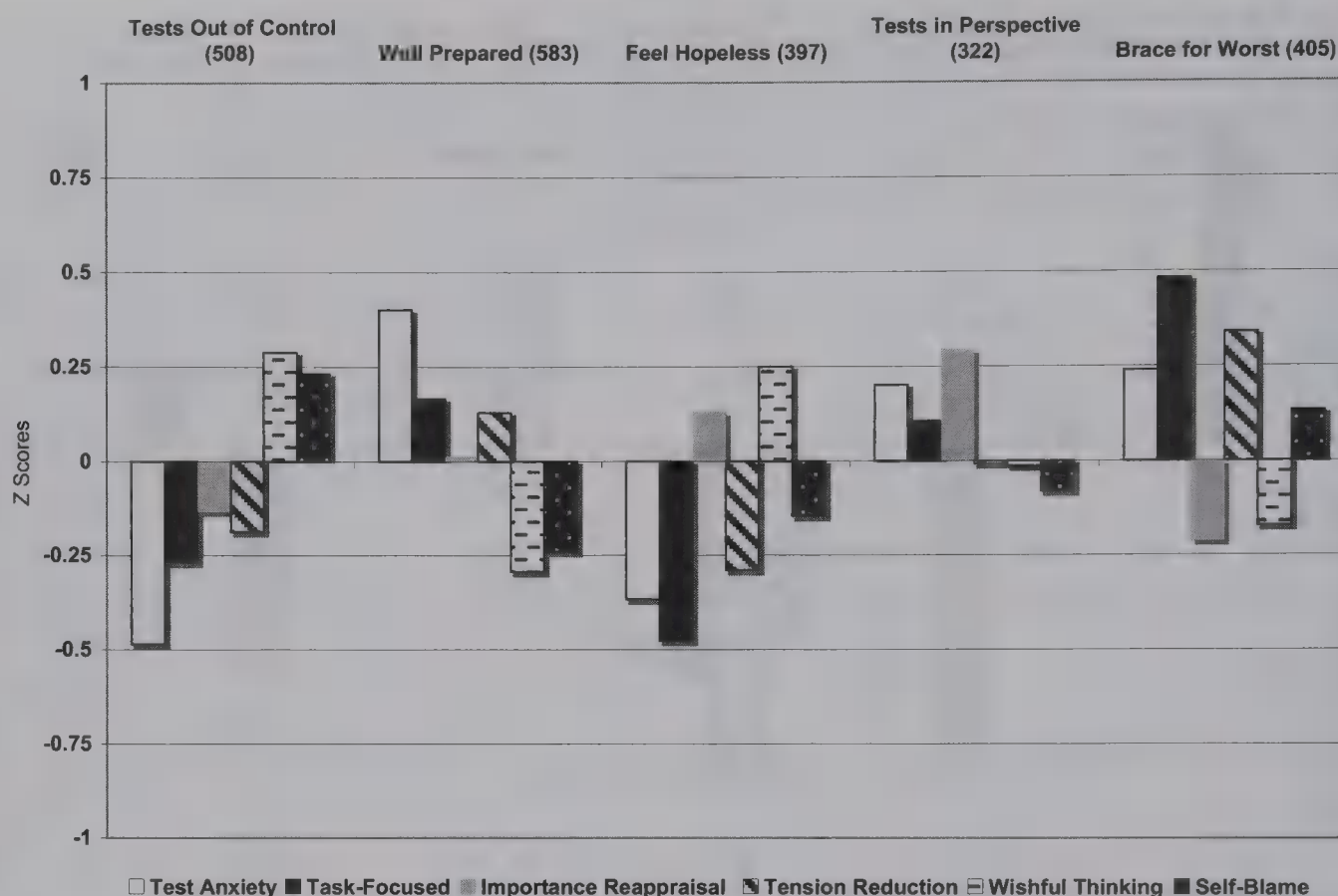


Figure 4. Patterns of anxiety and emotion-regulation strategies for the final, five-cluster sample. To assist with comparison, scores on the test anxiety and the five emotion-regulation strategy scales were converted to standardized scores. Note that higher scores on the Learning and Study Strategies Inventory Test Anxiety scale reflect lower rates of test anxiety.

these problems would be defined by the students' standards and goals. Thus, we speculate that that students in this cluster may not be achieving/performing to their standards.

*Cluster 4: Keeping tests in perspective.* This was, perhaps, the most challenging cluster for which to hypothesize differences. Because students in this cluster reported the lowest levels of goal importance and goal congruence, we hypothesized that students in this cluster would report some of the lowest levels of anxiety during tests. Consistent with appraisal theory, students in this cluster reported somewhat lower, but not the lowest, levels of anxiety during tests. However, because of lower reports of control (both agency and testing problem efficacy), we hypothesized that when faced with problems during tests, students would evidence a well-being pattern of regulation; that is, fewer reports of task-focusing regulation and higher reports of emotion-focusing strategies. Of interest, we found that students in this cluster deployed an unusual set of emotion-regulation processes, pairing moderate-to-high task-focusing strategies with some of the highest rates of importance reappraisal. Are these students oriented toward growth or well-being? Our data are a bit inconclusive. Although students in this cluster appear to have deployed strategies to manage the task of the test, they also reported that they concurrently chose to psychologically distance themselves by reappraising the importance of the test while taking the test. Although importance reappraisal during the test may have served to regulate the experience of anxiety, it may have some hidden consequences because they directed their attention away from what they could do to better

prepare for their next test. In future studies, researchers may want to examine the extent to which students in this cluster engage in procrastination strategies. In other words, to what extent are they deemphasizing the importance of the test prior to the testing situation, and how might that ultimately affect their test preparation practices (Kondo, 1997)? On the other hand, it may be that students in this cluster have different standards or goals that do not predispose them toward anxiety during tests. Although students in this cluster did not evidence the highest achievement, they also did not evidence the lowest achievement. Thus, perhaps students in this cluster merely go with the flow and avoid succumbing to the pressures of external evaluation.

*Cluster 5: Bracing for the worst.* Similar to Cluster 2, students in Cluster 5 evidenced patterns of regulation oriented toward growth (Boekaerts, 2007) and higher levels of achievement. We were uncertain about the extent to which students in this cluster would report experiencing anxiety. Of interest, students in this cluster reported lower, though not the lowest, levels of anxiety during tests. Moreover, they reported some of the lowest rates of distancing behaviors—that is, importance reappraisal and wishful thinking. When faced with problems during the test, these students deployed strategies designed to focus them on the problem of the test and their emotions (task focusing and tension reduction). These findings confirm all three theories: that when faced with problems during tests, and some anxiety, students activate strategies oriented toward growth. How do we then make sense of the finding that students in this cluster concurrently reported some of

the highest levels of self-blame? Similar to students in Cluster 2, we believe that this highly differentiated pattern of regulation, combined with students' appraisals and their higher scores on the SAT, suggest that problems during tests may actually have been somewhat rare. On the one hand, engaging in self-blame during tests may reflect a strategy of bracing for the worst (Shepperd, Findley-Klein, Kwavnick, Walker, & Perez, 2000; Taylor & Shepperd, 1998). Specifically, Taylor and Shepperd (1998) and Shepperd et al. (2000) argued that when managing emotions, students may engage in a tactic where they brace for the worst by blaming themselves. Whereas self-blame may be an emotion-focused strategy for students in Cluster 1, orienting them away from tests, self-blame may serve an activating, growth-oriented purpose for student in Cluster 5. This strategy may work for them because, in the end, the problems were minimal, and they experienced relief that they did better than they had anticipated. Moreover, their attention on what they struggled with during the test, may orient them toward making adaptive changes in their test preparation in the future. On the other hand, higher reports of self-blame may evidence a very low threshold for experiencing challenge during tests (Boekaerts, 2007). In the future, researchers may want to examine what happens when actual problems do arise during tests and students do feel anxious. We believe that students in this group may study hard and prepare for their exams but that they may tend to feel blindsided by difficulties on tests. When interpreting this cluster, we were reminded of a participant in an earlier interview project we completed (Davis et al., 1999).

For students like Jennifer, blaming herself and her preparation only worked to increase her anxiety (see also Kondo, 1997). "I lose my self-confidence by berating myself thinking, 'You know, I should have studied harder for the material. I should have looked it over more.' " This is because she already perceived herself as working very hard to prepare for the exam. Likewise, when I asked Marie what she wanted her professors to know about test anxiety, she pleaded, "Telling [students who feel anxious] to study harder does not make their problem *not* real." (Davis et al., 1999, p. 16)

Does the benefit of blaming oneself during the exam outweigh the cost of the anxiety it may produce? Though test anxiety may be less frequent, we wondered, for this group of students, whether anxiety may feel psychologically debilitating—even if faculty cannot see evidence of this debilitation in their actual achievement. Moreover, these students appear to be doing everything right, according to the appraisal and coping literature, but it may not be working for them. For students in this group, learning to put the test in perspective with their actual achievement, to engage in importance reappraisal during the test, may help them to manage anxiety that feels overwhelming.

### Implications for Theory and Practice

The purpose of this project was to explore whether there are different types of undergraduate test takers. Specifically, we were interested in examining patterns of students' appraisals about tests and patterns of coping with problems during tests. To accomplish this task, we surveyed a large sample of incoming undergraduates at a large Southwestern university and identified—across two independent samples—five unique groups, or clusters, of students that shared similar ways of approaching tests.

Findings from this study have important theoretical implications for all four of the theories of emotions in education described earlier. With regard to circumplex theory of emotions (Barrett, 2006; Linnenbrink, 2007; Russell, & Barrett, 1999), our findings may clarify why the experience of unpleasant, activating emotions like anxiety may not always lead to enhanced performance. Specifically, findings from Clusters 2, 4, and 5 reveal that when faced with problems and anxiety during tests, the students in these clusters may be activating different strategies, which may in turn have different implications for their performance. For example, when faced with problems during tests, students in Cluster 5 (bracing for the worst) reported a highly differentiated pattern of regulation drawing from task-focused processes, regaining task focus, and emotion-focused processes. Specifically, when students in this cluster activated task-focusing and self-blame strategies, it may have been to focus their attention toward elements of that test that they missed during preparation. In the short term, this strategy may have resulted in heightened anxiety during the test; however, this short-term cost may have been outweighed by the long-term benefit of increasing students' concentration and attention to detail and, in turn, may have contributed to their ability to modify preparation strategies and develop needed competencies. In contrast, students in Cluster 4 (putting tests in perspective) tended to activate a different set of strategies when faced with problems during tests. These students reported a high reliance on importance reappraisal, which may indicate that their attention was being drawn away from the test. Activation of different strategies may have caused these two groups' somewhat different patterns of achievement.

With regard to Pekrun et al.'s (2007) control-value theory of achievement emotions, our findings suggest that with two dimensions of value and two dimensions of control, we were able to consistently identify five different types of test takers, who experienced differential types of cognitive engagement with the test. Moreover, in line with Pekrun et al.'s theory, students who perceived themselves as having little control over the testing situation were more likely to report higher levels of emotion-focused coping, whereas students who perceived themselves as being in control and/or students who felt confident about managing problems were more likely to report higher levels of task-focused strategy use. Our findings, however, also add to this theory by noting the co-occurrence of task-focused and emotion-focused regulation processes within a single cluster. Future studies, perhaps with qualitative methodologies, should explore what function the constellation of strategies serves for each type of test taker as well as the costs and benefits associated with each strategy (Thompson & Calkins, 1996).

With regard to Boekaert's (1993, 2007) dual-process theory, our findings appear to confirm that there are two self-regulatory pathways students may assume while taking tests. For students in Clusters 1, 3, and 4 (out of control, feeling helpless, test in perspective), problems on tests appear to have signaled to them that something was wrong, and in lieu of choosing strategies that might have refocused them on the learning goals, these students reported deploying strategies that psychologically distanced them from the test. Although these students reported that they experienced less anxiety during tests, they also appear to have engaged in strategies that may have benefited them in the short-term by preserving their sense of well-being, but had long-term conse-



quences in that they may have distanced the students from the goal of the assessment. It is important to remember that these data were drawn from students entering their first semester at college. What might be the long-term implications of using distancing strategies? Future studies may want to explore the longitudinal impact of these orientations toward test taking.

Findings from this study also have important theoretical implications for considering both the dynamic interplay of cognitive appraisals and emotion-regulation strategies (Schutz & Davis, 2000; Schutz & DeCuir, 2002; Schutz et al., 2004). First, our findings point to the complexity of the emotion-appraisal process. Traditional appraisal theory would argue that students who do not judge tests to be important should not experience anxiety during tests. This is because the primary appraisal of importance is viewed as a necessary condition for emotion to occur. However, our data suggest that students in the feeling hopeless cluster, despite reporting that tests are not important, tended to report experiencing higher levels of anxiety during test taking. These findings corroborate Boekaerts's (2007) model, which states that perceived alignment with students' goals (i.e., goal congruence) is as important as goal importance. Additionally, we wondered about the extent to which students truly see a distinction between goal relevance and goal congruence. Although appraisal theory would suggest that these two appraisals are distinct, our findings suggest that, in general, students who reported that tests were relevant to their goals also tended to report that their test outcomes were goal congruent (see Table 4). In future studies, researchers may want to examine the extent to which, within the test-taking context, these appraisals of tests are truly unique.

Second, our findings point to the complex nature of coping strategies—specifically how one set of strategies viewed as unfavorable in the literature (i.e., self-blame) might work well for a particular group of students (e.g., those in the tests in perspective cluster). This strategy is conceptualized as emotion focused and, therefore, as a less adaptive coping strategy, because it draws students' attention and emotional resources away from the task at hand. In future studies, researchers may want to explore how students use self-blame as a way of focusing their attention (Gross & Munoz, 1995; Gross & Thompson, 2007). Could it be that self-blame means different things for different types of students? Indeed, findings from our earlier qualitative work (Davis et al., 1999) suggested that some high-achieving students tended to use self-blame in a more task-focusing way. In other words, blaming themselves for missing an item helped them to refocus their efforts for the current as well as the next test. These students' language and use of self-blame challenges the ways in which researchers have traditionally conceptualized that form of coping. Thus, additional qualitative work, attempting to identify students from each cluster for interviewing, might help researchers understand the ways in which students understand different strategies for coping with tests.

Our findings could make a significant practical contribution to the field of undergraduate student services (Tuckman, 2007). Our findings suggest that there is no single type of test taker—and that helping students to learn to manage test taking may require a more personalized approach. We hope our findings provide some practical information for faculty and instructors of undergraduate students, which can be used to help students with less adaptive

patterns of appraisals or emotion-regulation strategies to modify their approach to taking tests. Moreover, our findings may assist faculty and instructors in tailoring their feedback to fit the students they are interacting with. We hope, in particular, that our findings will help faculty and instructors to hear students when they say they struggle with anxiety during tests. For some instructors, this may mean noting that counseling students on the importance of test taking, test prep, and the consequences of test scores and grades would be an ineffective strategy for several of our clusters who may already overemphasize the importance of test scores (i.e., bracing for the worst). Moreover, educating students to engage in task-focusing and regaining-task-focusing strategies may cause some students to feel frustrated and misunderstood (e.g., students in the bracing for the worst cluster, who are already engaging in high levels of task-focused and regaining task-focused behavior). Instead, our data suggest that students' prior beliefs about test taking need to first be brought to the forefront. Some students may need to re-evaluate the extent to which tests are important to their lives, which may help students who undervalue tests (test in perspective cluster) and those who overvalue tests (bracing for the worst cluster) to consider the role that tests really play in their life goals. Moreover, careful counseling in course selection, to match challenge level with ability and to identify instructors who will work closely with students, may be the necessary first step in creating opportunities for mastery experiences for students who feel hopeless.

It is important to keep in mind that the study of emotion in educational settings is challenging. For this study, we used cross-sectional, self-report data collected to measure trait-type emotion and emotional regulation processes. In other words, in collecting the data, we asked our participants to think about testing and what they do during the testing process in general; therefore, our data were not tied to a specific test. Thus, although the data provide us with valuable insight into what might be considered trait-type processes during testing, research tied to a particular test (e.g., state emotion and emotional regulation), research that is more longitudinal in nature (e.g., following how emotions emerge and potentially change during the course of a particular test), and research that uses data-collection methods that potentially go beyond self-report methods (e.g., galvanic skin response) should be conducted.

As such, future research should examine the extent to which our five-cluster solution replicates in other populations of college students. Given the large sample size and the high hit rates across samples, we feel confident that these clusters are likely to appear in other samples. However, college students represent a relatively academically homogeneous population (those with high enough achievement to be admitted into college) compared with the larger population of students in public high schools. Thus, future studies should examine whether additional approaches to taking tests are used by lower achieving or more academically diverse populations of students. Likewise, future studies should examine the stability, throughout students' academic careers, and the validity, in other populations of students, of the five-cluster solution. Although we were successful in identifying the clusters across two independent samples, our challenges in having a clean cross-classification lead us to be hesitant to talk about the utility of the ERT as a diagnostic tool. On the other hand, challenges with cross-classification could have been an artifact of sample-specific characteristics. Future



studies should further examine the external validity of the solution, including examining whether the five-cluster solution can be distinguished with external indicators, such as first-year college students' expectations about college, their learning strategies, and their procrastination strategies (Tuckman, 2005) used in high school. This may help to illuminate whether the ERT could be scored, with our prediction rule, by an academic counselor to help classify students who may be academically at risk and to tailor interventions designed to both reduce anxiety and improve performance. Finally, it is important to remember that researchers studying emotion regulation have argued that task-focused strategies refer not only to the deployment of attention, but also to qualitatively different ways of processing information and solving problems. In future studies, researchers may want to elaborate on Schutz and colleagues' (Schutz & Davis, 2000; Schutz et al., 2004, in press) concept of task-focused processes to examine the ways in which different clusters of students solve the problem of tests in qualitatively different ways.

## References

- Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis*. Beverly Hills, CA: Sage.
- Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic Press.
- Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, 84, 191–215.
- Barrett, L. F. (2006). Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, 10, 20–46.
- Ben-Ze'ev, A. (2000). *The subtlety of emotion*. Cambridge, MA: MIT Press.
- Blashfield, R. K., & Aldenderfer, M. S. (1988). The methods and problems of cluster analysis. In J. R. Nesselrode & R. B. Cattell (Eds.), *International handbook of multivariate experimental psychology* (pp. 130–174). New York: Plenum Press.
- Boekaerts, M. (1993). Being concerned with well-being and with learning. *Educational Psychologist*, 28, 149–167.
- Boekaerts, M. (2007). Understanding students' affective processes in the classroom. In P. A. Schutz & R. Pekrun (Eds.), *Emotion in education* (pp. 37–56). Amsterdam: Elsevier.
- Brdar, I., Rijavec, M., & Loncaric, D. (2006). Goal orientations, coping with school failure and school achievement. *European Journal of Psychology of Education*, 21, 53–70.
- Carver, C. S., & Scheier, M. F. (2000). On the structure of behavioral self-regulation. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 41–84). San Diego, CA: Academic.
- Carver, C. S., Scheier, M. F., & Weintraub, J. K. (1989). Assessing coping strategies: A theoretically based approach. *Journal of Personality and Social Psychology*, 56, 267–283.
- Davis, H. A., DiStefano, C., DeCuir, J. T., & Schutz, P. A. (2000, April). *Patterns of appraisal and emotion regulation during test taking*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Davis, H. A., Schutz, P. A., & DeCuir, J. T. (1999, August). "They can't tell me I'm stupid": Undergraduate students coping with test anxiety. Paper presented at the annual meeting of the American Psychological Association, Boston, MA.
- DeCuir, J. T., Aultman, L. P., & Schutz, P. A. (in press). Investigating transactions among approach/avoidance motives, emotions and emotional regulation during testing. *Journal of Experimental Education*.
- DiStefano, C., & Kamphaus, R. W. (2006). Investigating subtypes of child development. *Educational & Psychological Measurement*, 66, 778–794.
- DiStefano, C., Kamphaus, R. W., Horne, A., & Windsor, A. (2003). Behavioral adjustment in the U.S. elementary school: Cross-validation of a person-oriented typology of risk. *Journal of Psychoeducational Assessment*, 21, 338–357.
- Ellsworth, P. C., & Smith, C. A. (1988). From appraisal to emotion: Differences among unpleasant feelings. *Motivation & Emotion*, 12, 271–302.
- Folkman, S., & Lazarus, R. S. (1985). If it changes it must be a process: Study of emotion and coping during three stages of a college examination. *Journal of Personality and Social Psychology*, 48, 150–170.
- Gross, J. J. (1998). Antecedent- and response-focused emotion regulation: Divergent consequences for experience, expression, and physiology. *Journal of Personality and Social Psychology*, 74, 224–237.
- Gross, J. J. (1999, September). Emotion regulation: Past, present, future. *Cognition & Emotion*, 13, 551–573.
- Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, 39, 281–291.
- Gross, J. J., & Munoz, R. F. (1995). Emotion regulation and mental health. *Clinical Psychology: Science and Practice*, 2, 151–164.
- Gross, J. J., & Thompson, R. A. (2007). Emotional regulation; Conceptual foundations. In J. J. Gross (Ed.), *Handbook of emotion regulation* (pp. 3–26). New York: Guilford Press.
- Hodapp, V., & Benson, J. (1997). The multidimensionality of test anxiety: A test of different models. *Anxiety, Stress, and Coping*, 10, 219–244.
- Hembree, R. (1998). Correlates, causes, effects and treatment of test anxiety. *Review of Educational Research*, 58, 7–77.
- Huberty, C. J. (1994). *Applied discriminant analysis*. New York: Wiley.
- Huberty, C. J., DiStefano, C., & Kamphaus, R. W. (1997). Behavioral clustering of school children. *Multivariate Behavioral Research*, 32, 105–134.
- John, O. P., & Gross, J. J. (2004). Healthy and unhealthy emotion regulation: Personality processes, individual differences, and lifespan development. *Journal of Personality*, 72, 1301–1334.
- John, O. P., & Gross, J. J. (2007). Individual differences in emotion regulation. In J. J. Gross (Ed.), *Handbook of emotion regulation* (pp. 351–372). New York: Guildford Press.
- Kondo, D. S. (1997). Strategies for coping with test anxiety. *Anxiety, Stress, and Coping*, 10, 203–215.
- Lazarus, R. S. (1985). The costs and benefits of denial. In A. Monat & R. S. Lazarus (Eds.), *Stress and coping: An anthology* (2nd ed., pp. 154–173). New York: Columbia University Press.
- Lazarus, R. S. (1991). *Emotion and adaptation*. New York: Oxford University Press.
- Lazarus, R. S. (1999). *Stress and emotions: A new synthesis*. New York: Springer.
- Lazarus, R. S. (2001). Relational meanings and discrete emotions. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion* (pp. 37–67). New York: Oxford University Press.
- Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. New York: Springer.
- Lazarus, R. S., & Folkman, S. (1991). The concept of coping. In A. Monat & R. S. Lazarus (Eds.), *Stress and coping: An anthology* (3rd ed., pp. 189–206). New York: Columbia University Press.
- Linnenbrink, L. A. (2007). The role of affect in student learning: A multi-dimensional approach to considering the interaction of affect, motivation, and engagement. In P. A. Schutz & R. Pekrun (Eds.), *Emotion in education* (pp. 13–36). Amsterdam: Elsevier.
- Linnenbrink, L. A., & Pintrich, P. R. (2002). The role of motivational beliefs in conceptual change. In M. Limon & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 115–135). Dordrech, the Netherlands: Kluwer Academic.
- Linnenbrink, L. A., Ryan, A. M., & Pintrich, P. R. (1999). The role of goals and affect in working memory functioning. *Learning and Individual Differences*, 11, 213–220.



- Martin, R. C., & Dahlen, E. R. (2005). Cognitive emotion regulation in the prediction of depression, anxiety, stress, and anger. *Personality and Individual Differences*, 39, 1249–1260.
- Milligan, G. W., & Cooper, M. C. (1987). Methodology review: Clustering methods. *Applied Psychological Measurement*, 11, 329–354.
- Moos, R. H., & Holahan, C. J. (2003). Dispositional and contextual perspectives on coping: Toward an integrative framework. *Journal of Clinical Psychology*, 59, 1387–1403.
- Panksepp, J. (2005). Affective consciousness: Core emotional feelings in animals and humans. *Consciousness and Cognition*, 14, 30–80.
- Pekrun, R. (2000). A social cognitive, control-value theory of achievement emotions. In J. Heckhausen (Ed.), *Motivational psychology of human development* (pp. 143–163). Oxford, England: Elsevier Science.
- Pekrun, R., Frenzel, A. C., Goetz, T., & Perry, R. (2007). The control-value theory of achievement emotions: An integrative approach to emotions in education. In P. A. Schutz & R. Pekrun (Eds.), *Emotion in education* (pp. 13–36). Amsterdam: Elsevier.
- Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, 37, 91–105.
- Rafnsson, F. D., Jonsson, F. H., & Windle, M. (2006). Coping strategies, stressful life events, problem behaviors and depressed affect. *Anxiety, Stress, and Coping*, 19, 241–257.
- Rice, K. G., & Slaney, R. B. (2002). Clusters of perfectionists: Two studies of emotional adjustment and academic achievement. *Measurement and Evaluation in Counseling and Development*, 35, 35–48.
- Roseman, I., Anoniou, A. A., & Jose, P. (1996). Appraisal determinants of emotions: Constructing a more accurate and comprehensive theory. *Cognition and Emotion*, 10, 241–277.
- Roseman, I., & Smith, C. A. (2001). Appraisal theory: Overview, assumptions, varieties, and controversies. In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion* (pp. 3–19). New York: Oxford University Press.
- Rosenberg, E. L. (1998). Levels of analysis and the organization of affect. *Review of General Psychology*, 2, 247–270.
- Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, 76, 805–819.
- Sarason, I. G., & Sarason, B. R. (1990). Test anxiety. In H. Leitenberger (Ed.), *Handbook of social and evaluation anxiety* (pp. 475–495). New York: Plenum.
- Scherer, K. R. (1988). Criteria for emotion-antecedent appraisal: A review. In V. Hamilton, G. H. Bower, & N. H. Frijda (Eds.), *Cognitive perspectives on emotion and motivation* (pp. 89–126). Norwell, MA: Kluwer Academic.
- Schutz, P. A., Benson, J., & DeCuir, J. T. (in press). Approach/avoidance motives, test emotions, and emotional regulation during testing. *Anxiety, Stress and Coping: An International Journal*.
- Schutz, P. A., & Davis, H. A. (2000). Emotions during self-regulation: The regulation of emotions during test taking. *Educational Psychologist*, 35, 243–256.
- Schutz, P. A., & DeCuir, J. T. (2002). Inquiry on emotions in education. *Educational Psychologist*, 37, 125–134.
- Schutz, P. A., DiStefano, C., Benson, J., & Davis, H. A. (2004). The Emotional Regulation During Test Taking Scale. *Anxiety, Stress, and Coping*, 17, 253–259.
- Schutz, P. A., & Pekrun, R. (2007). *Emotion in education*. Amsterdam: Elsevier.
- Shepperd, J. A., Findley-Klein, C., Kwavnick, K. D., Walker, D., & Perez, S. (2000). Bracing for loss. *Journal of Personality and Social Psychology*, 78, 620–634.
- Smith, C. A. (1991). The self, appraisal, and coping. In C. R. Snyder & D. R. Forsyth (Eds.), *Handbook of social and clinical psychology: The health perspective* (pp. 116–137). Elmsford, NY: Pergamon.
- Smith, C. A., & Ellsworth, P. C. (1987). Patterns of appraisal and emotions related to taking exams. *Journal of Personality and Social Psychology*, 52, 475–488.
- Smith, C. A., Haynes, K. N., Lazarus, R. S., & Pope, L. K. (1993). In search of the "hot" cognitions: Attributions, appraisals, and their relation to emotion. *Journal of Personality and Social Psychology*, 65, 916–929.
- Sondaite, J., & Zukauskienė, R. (2005). Adolescents' social strategies: Patterns and correlates. *Scandinavian Journal of Psychology*, 46, 367–374.
- Tanaka, K. (2007). Relations between general goal orientations and task-specific self-appraisals. *Japanese Psychological Research*, 49, 235–247.
- Taylor, K. M., & Shepperd, J. A. (1998). Bracing for the worst: Severity, testing and feedback as moderators of the optimistic bias. *Personality and Social Psychology Bulletin*, 24, 915–926.
- Thompson, R. A., & Calkins, S. D. (1996). The double-edged sword: Emotional regulation for children at risk. *Development and Psychopathology*, 8, 163–182.
- Tuckman, B. W. (2005). Relations of academic procrastination on rationalizations and performance in a Web course with deadlines. *Psychological Reports*, 96, 1015–1021.
- Tuckman, B. W. (2007, April). *Evaluating a program for enhancing the study skills and academic performance of urban high school students*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- van Reekum, C. M., Johnstone, T., Banse, R., Etter, A., Wehlre, T., & Scherer, K. R. (2004). Psychophysiological responses to appraisal dimensions in a computer game. *Cognition and Emotion*, 18, 663–688.
- Weinstein, C. E., & Palmer, D. R. (2002). *User's manual: Learning and Study Strategies Inventory* (2nd ed.). Clearwater, FL: H & H.
- Weinstein, C. E., Palmer, D. R., & Hanson, G. R. (1995). *Perceptions, expectations, emotions, and knowledge about college*. Clearwater, FL: H & H.
- Weinstein, C. E., Palmer, D. R., & Schulte, A. C. (1987). *Learning and Study Strategies Inventory*. Clearwater, FL: H & H.
- Zeidner, M. (1995). Coping with examination stress: Resources, strategies, and outcomes. *Anxiety, Stress, and Coping*, 8, 279–298.
- Zeidner, M. (1996). How do high school and college students cope with test situations? *British Journal of Educational Psychology*, 66, 115–128.
- Zeidner, M. (1998). *Test anxiety: The state of the art*. New York: Plenum.
- Zeidner, M. (2007). Test anxiety in educational contexts: Concepts, findings, and future directions. In P. A. Schutz & R. Pekrun (Eds.), *Emotion in education* (pp. 13–36). Amsterdam: Elsevier.
- Zeidner, M., & Matthews, G. (2005). Evaluative anxiety. In A. Elliott & C. Dweck (Eds.), *Handbook of competence and motivation* (pp. 141–166). New York: Guilford Press.

Received March 5, 2007

Revision received May 15, 2008

Accepted May 17, 2008 ■

# Confidence and Cognitive Test Performance

Lazar Stankov and Jihyun Lee  
Educational Testing Service

This article examines the nature of confidence in relation to abilities, personality, and metacognition. Confidence scores were collected during the administration of Reading and Listening sections of the Test of English as a Foreign Language Internet-Based Test (TOEFL iBT) to 824 native speakers of English. Those confidence scores were correlated with performance accuracy scores from the TOEFL iBT and SAT, high school grade point averages (HS-GPA), and measures of personality and metacognition. The results of factor analyses indicate that confidence is a separate psychological trait, somewhere between ability and personality. The findings also suggest that confidence is related to, but separate from, metacognition. Gender and ethnic differences in confidence are also reported, with men and African Americans showing higher overconfidence bias than women and Whites or Hispanics, respectively. Finally, the data show small incremental validity of the confidence scores above and beyond the accuracy scores in predicting performance on the TOEFL iBT as a whole, the Writing and Speaking sections of the TOEFL iBT, and a test of numeracy. Confidence does not show incremental validity for the SAT and HS-GPA.

**Keywords:** confidence, overconfidence bias, metacognition, self-monitoring

Psychologists in the field of decision-making and education inspired the work on confidence by questioning whether those who know more also know more about how much they know. *Know* refers to performance accuracy, and *knowing how much they know* relates to confidence (Lichtenstein & Fischhoff, 1977). Two important theoretical approaches have been dominant in the study of confidence within the decision-making tradition: the ecological approach (Gigerenzer, Hoffrage, & Kleinbolting, 1991) and the heuristics and biases approach (Kahneman, Slovic, & Tversky, 1982). According to the ecological approach, the discrepancy between what we know and knowing how much we know can be explained in terms of the choice of difficult and “tricky” items for use in cognitive tests. These are the items for which a person does not know the correct answer and chooses to use the available cues to answer the question. These cues may be misleading. Thus, for the advocates of the ecological (sometimes also referred to as Brunswikian; see Juslin & Olsson, 1997) position, the sources of miscalibration reside outside the individual.

The heuristics and biases approach, on the other hand, attributes miscalibration to the sources from within the individual. This latter approach is sometimes called Thurstonian and is linked to the “error” model proposed by Soll (1996). According to the error model, miscalibration is due to participants’ inconsistency in forming subjective feelings of confidence and/or limited experience with the environment. This random error interacts with stimulus structure and may lead to overconfidence. Dougherty (2001) pro-

posed an integration of the ecological and error models using a modification of Hintzman’s (1988) multiple-trace memory model. This latter model predicts that miscalibration (i.e., overconfidence bias) should decrease both as a function of experience and as a function of encoding quality. Assuming that people with higher cognitive ability have more experience and better encoding quality, Dougherty’s (2001) model can be interpreted to mean that people with higher cognitive ability (or IQ) will show lower overconfidence bias.

Within the area of education, the same question is asked under the rubric of metacognition (Kleitman & Stankov, 2001; Schraw & Dennison, 1994; Stankov, 2000; Tobias & Everson, 2000, 2002). The approach of the present article has its roots in educational testing tradition, and theoretically it is Thurstonian in spirit. In our view, internally located sources of error consist of a random part that is suggested by Soll (1996) and systematic and reliable, i.e., nonrandom, individual differences in confidence. This confidence trait can play an important part in the explanations of miscalibration.

Our primary focus in this article is on the nature of confidence and to a lesser degree on the discrepancy between measures of confidence and accuracy of performance. Our aim is to demonstrate that confidence can be measured reliably and that it is a psychological trait, conceptually similar to personality and ability traits. At the group level of analysis, we examine gender and ethnic differences in performance accuracy, confidence, and overconfidence bias to highlight the potential usefulness of these measures for our understanding of behavior. The primary emphasis, however, is on individual differences in confidence, because, as we point out in a later section of this article, there are concerns about the measurement properties of overconfidence bias scores.

## Individual Differences in Confidence

Interest in confidence has a long history in psychology. Psychological studies of confidence started with the work of Fullerton

---

Lazar Stankov and Jihyun Lee, Center for New Constructs, Educational Testing Service, Princeton, New Jersey.

Correspondence concerning this article should be addressed to Lazar Stankov, Center for New Constructs, 16R, Educational Testing Service, Rosedale Road, Princeton, NJ 08541. E-mail: lstankov@ets.org or jlee@ets.org



and Cattell (1882). Trow (1923) and Festinger (1943a, 1943b), among others, reported further significant work. Classical psychophysicists routinely collected three bits of information in their studies of threshold performance: accuracy, speed, and confidence. More recent studies in this area were reviewed by Vickers (1979) and Baranski and Petrusic (1999). Confidence in psychophysics is often linked to speed of providing an answer, but more complex cognitive tasks point to a relative independence of confidence and speed measures (see Stankov, 2000).

Much of the work in both decision-making and psychophysics has been experimental rather than differential in nature. It was only in the late 1990s that systematic individual differences were seriously considered as more than a part of error in the decision-making process. There is already a substantial amount of data showing individual differences in confidence ratings (Pallier et al., 2002; Schraw & Dennison, 1994; Stankov, 1998, 1999; Stankov & Crawford, 1996, 1997; Stanovich, 1999). From the individual differences point of view, confidence is a disposition (i.e., a systematic tendency that leads one to act in a particular way because it reflects a belief, a faith in oneself). Those scoring high on it can be described as decisive, firm, and resolute. Those scoring low, on the other hand, are doubtful of their capacity. Stankov (1999) has placed confidence in the no man's land between personality and abilities. Along with other dispositional measures, such as those of typical intellectual engagement, various self constructs, outlooks, and perhaps emotional intelligence, the essence of confidence cannot be reduced to either ability or personality constructs. Its advantage over most of these other constructs is the presence of a reality check in that confidence can be compared to the actual cognitive performance. This reality check operates at the subjective level (How confident am I that my answer is correct?), and, as this article discusses later, it can be assessed with an objective method (e.g., over a series of items in a test).

### The Measurement of Confidence

Accumulated evidence shows that confidence can be reliably measured in typical test-taking situations (see Crawford & Stankov, 1996). The procedure employed in the present study for assessing confidence follows Crawford and Stankov's (1996) approach. Participants are asked to give a rating (expressed in terms of percentages) immediately after responding to an item in a test, indicating how confident they are that the chosen answer for this item is correct (see Crawford & Stankov, 1996; Harvey, 1997; Keren, 1991; Stankov, 1999). Thus, these ratings directly follow the cognitive act of providing an answer. Confidence ratings for all attempted test items are averaged to give an overall confidence score. Confidence scores can be studied on their own. They can also be studied in relationship to the performance accuracy measures obtained from the same cognitive test (Kleitman & Stankov, 2001; Stankov, 2000). These are sometimes referred to as calibration studies.

In studying the relationship between confidence and accuracy on a typical cognitive test, performance accuracy scores are often converted to average percentage correct and subtracted from the average percentage confidence scores. The result is called the *bias* (or *realism*) score. There is a considerable amount of data showing overconfidence bias in cognitive tasks—that is, a pronounced

(positive) difference between confidence and accuracy scores (see Lichtenstein & Fischhoff, 1977; Lichtenstein, Fischhoff, & Phillips, 1982). Pronounced underconfidence bias has been found with some sensory tasks (Olsson & Winman, 1996). However, we use the term bias interchangeably with overconfidence bias in the present study, because all tasks employed in this study are known to display overconfidence (Stankov, 2000). Zero is often treated as an ideal value for bias scores because a good match, frequently referred to as good calibration, between the level of confidence and performance is often seen as a desirable characteristic. In this article, we do not report bias scores at the individual levels of analysis. We report only arithmetic means of bias scores (i.e., differences between the means for accuracy and confidence scores) to point out the discrepancies between subpopulations.

### Correlates of Confidence

Recent studies show that confidence is a general trait and structurally independent of other established ability and personality traits (see Blais, Thompson, & Baranski, 2005; Klayman, Soll, Gonzalez-Vallejo, & Barlas, 1999; Kleitman & Stankov, 2001; Pallier et al., 2002; Stankov, 1998, 1999, 2000; Stankov & Crawford, 1996, 1997). There are three main domains that have been conceptually and empirically related to the construct of confidence.

First, confidence scores have shown moderate correlations with measures of *cognitive abilities*—that is, a higher accuracy score is linked to a higher confidence level. The list of different cognitive tasks employed in studies of confidence to date includes measures of verbal and nonverbal reasoning (fluid intelligence, *Gf*); acculturated knowledge (crystallized intelligence, *Gc*); long-term and short-term memory; and visual, olfactory, gustatory, and auditory perceptual processes, among others. In general, the studies show that people who are more confident on one cognitive task tend to be more confident across other tasks. Although evidence indicates that correlations between performance accuracy and confidence scores from the same tests tend to average between .40 and .60 (Stankov, 2000), correlations between confidence scores from different cognitive tests have been equal to or higher than correlations between performance accuracy and confidence scores from the same tests. That suggests a general confidence factor that is separate from (yet positively related to) factors of intelligence. Confidence as a separate factor has been identified in studies of Kleitman and Stankov (2001); Pallier et al. (2002); Stankov (1998, 1999), and Stankov and Crawford (1996, 1997; see also Stanovich, 1999).

Second, confidence is sometimes treated as a *personality trait*, either on its own or as an underlying facet of broader traits (Blais, Thompson, & Baranski, 2005; Pallier et al., 2002). Consistent small correlations ( $r = .30$ ) between confidence and the openness factor from the Big Five model have been noted (Pallier et al., 2002). However, it is unclear whether this relationship with openness is mediated by cognitive ability, because both confidence and openness correlate with cognitive ability. Moreover, Blais et al. (2005) demonstrated that a broad range of cognitive styles, including the need for cognition and the desire for structure, had no effect on confidence. Conceptually, cognitive styles straddle the boundary between personality and metacognition.



Third, as mentioned in the introduction, confidence judgments are frequently interpreted as an important aspect of *metacognitive* processes. Metacognition refers to an awareness of the learning process and one's cognitive strengths and weaknesses and therefore affects confidence judgments. In the writings of Nelson and Narens (1992, 1994) on metacognition, confidence ratings are assigned a somewhat limited role—they are assumed to provide no other metainformation apart from whether to retrieve an answer or continue a cognitive search further. Consequently, these researchers are mainly concerned with investigations of the ease of learning, judgments of learning, and feeling of knowing, often neglecting unique confidence ratings. On the other hand, many researchers see confidence judgments as noteworthy topics for psychological study. For instance, Baron (2000) argued that “appropriate confidence is, in most cases, a more realistic goal than certainty” (p. 62). That is, the metacognitive ability to appraise the relevance of one's own beliefs about the available evidence (and one's own performance) is an important factor in effective decision-making. One component of metacognition, self-monitoring, refers to the awareness of the accuracy of one's answers on typical cognitive tests. Bias score can be used as a direct measure of self-monitoring (see Schraw & Dennison, 1994). Broader aspects of metacognition are frequently assessed with questionnaires. Although questionnaire measures of metacognition have shown a moderate relationship with confidence in studies by Kleitman (2003) and Kleitman and Stankov (2006), those measures also defined a separate metacognitive factor that was distinct from the confidence factor itself. In this article we ask the same question—that is, do questionnaire measures of metacognition define a different factor from measures of confidence?

### Gender and Ethnic Differences in Confidence and Bias Scores

Gender differences in confidence and bias scores have been examined extensively, but the findings are less than conclusive so far. Studies by Stankov and Crawford (1996, 1997) found no gender differences, but more recent studies by Pallier et al. (2002) and Pallier (2003) show significant gender differences in confidence, with women exhibiting lower bias scores (i.e., better calibration) on cognitive tasks. We address the gender difference issue in this article.

The evidence for possible ethnic differences in confidence has been explored only indirectly so far. Stankov and Kim (2006) found that on tests of esoteric analogies and vocabulary, foreign students seeking admission to U.S. universities tend to have overconfidence bias scores that are about twice as high as U.S. students. At present, there is no available information on whether different ethnic groups within a country show pronounced differences in confidence ratings. In the present study, we examine ethnic differences in confidence and bias scores between Whites, Hispanics, and African American participants.

### Evidence for Incremental Validity of Confidence Scores

The evidence for incremental validity of confidence scores is not available at present. The concept of incremental validity is based on the question of whether a measure can predict a criterion over and above what has been predicted by some other variable. In

other words, can confidence predict some sort of educational and job-related outcome after taking account of the accuracy of performance on the test? The focus, to date, has been on the understanding of confidence scores themselves (Kleitman & Stankov, 2006). In this study, we address the issue of their predictive validity using various measures of aptitude as criteria (i.e., self-reported SAT scores, high school grade point average [HS-GPA], numeracy test scores, Test of English as a Foreign Language [TOEFL] scores for the Writing and Speaking sections). These are all measures that are known to predict both school achievement and training and job-related performance.

Our approach to measuring confidence is to obtain accuracy and confidence scores from the same cognitive act. In showing the incremental validity of confidence scores above and beyond their yoked accuracy scores, we predict accuracy scores from cognitive measures that are not used to extract confidence scores. For example, the question is whether confidence scores that are obtained from the Reading sections in the TOEFL exam will add to the prediction of, say, HS-GPA above and beyond TOEFL Reading performance accuracy scores.<sup>1</sup> This is a stringent requirement, because performance accuracy scores from both criterion and predictor sets of measures (HS-GPA and TOEFL Reading section) are capturing the same construct—acculturated knowledge (or crystallized ability, *Gc*). The presence of incremental validity may contribute to the argument for using confidence scores in selection and guidance programs.

Schmidt and Hunter (1998) showed that only two noncognitive constructs—conscientiousness and integrity—have incremental validity beyond performance accuracy scores. In this context, even a small amount of incremental validity that can be attributed to confidence after taking account of measures of performance accuracy from the same test will be of importance.

### TOEFL Internet-Based Test (iBT) as a Measure of Verbal Comprehension Ability in Native Speakers of English

We assess confidence during the administration of a TOEFL iBT exam to native speakers of English. Although confidence expressed in each TOEFL item may be confined to participants' confidence in his or her command of English, confidence is a part of any cognitive activity that requires decision-making, not just foreign-language learning. Thus, in this study, confidence scores based on subjective judgments of the accuracy of each item in the TOEFL exam are used in the same way as any other cognitive test can be used.

The most recent version of the TOEFL exam (known as the TOEFL iBT) consists of four sections: Reading, Listening, Writing, and Speaking. The test is delivered via the Internet. Detailed analyses of the TOEFL iBT exam based on nonnative speakers of English are provided in the Educational Testing Service's technical report by Sawaki, Stricker, and Oranje (2005). As has been the case with previous versions of TOEFL, the validation process includes administration of the test to native speakers of English (see Angelis, Swinton, & Cowell, 1979; Angoff & Sharon, 1971; Johnson, 1977; Stricker, 2002). The present study is based on such

<sup>1</sup> This is just an example. We do not expect to predict HS-GPA particularly well by aspects of language proficiency captured by the TOEFL iBT.



a sample of native speakers from community colleges who took the TOEFL iBT exam. There is considerable evidence showing that TOEFL performance correlates with other cognitive tests (Stricker, 2002). With nonnative speakers of English, the Graduate Record Examination verbal reasoning score correlation is .61 with the TOEFL total score. For native speakers of English, this correlation is .64 (Stricker, 2002). This finding has led to a speculation that it may be possible to use the TOEFL exam as a proxy for the Graduate Record Examination or SAT tests under some circumstances.<sup>2</sup> Within a broad area of human cognitive abilities, TOEFL measures verbal comprehension primary ability and acculturated knowledge (*Gc*) at the second-order level.

Two sections in the TOEFL iBT exam—Reading and Listening—contain multiple-choice items, the item type that has been employed in most previous studies of confidence. A sample of these TOEFL multiple-choice items was used to collect confidence ratings in the present study. Performance accuracy scores on the TOEFL Speaking and Listening sections as well as the TOEFL total scores were also used in our study (see the Method section in this article). To check the interpretation of TOEFL accuracy scores as measures of cognitive ability and *Gc* in particular, we included in this study a numeracy test from the literacy battery developed by Kirsch et al. (2001), as well as an overclaiming test—a self-reported word knowledge measure employed in studies of overclaiming (see Paulhus & Harms, 2004).

### Goals

The purpose of this study is to examine the nature of confidence exhibited during performance on two sections (Reading and Listening) of a cognitive test, the TOEFL iBT assessment. Two issues are addressed.

First, we examine psychometric properties of confidence scores that are obtained from TOEFL iBT Reading and Listening sections—their reliability and validity. In accordance with the findings from previous studies that employed different cognitive tests, the expectation is that confidence scores from TOEFL iBT Reading and Listening tests show satisfactory reliabilities (test–retest, parallel forms, and internal consistency). Two broad classes of validity evidence for confidence scores are considered in this article.

*Construct validity* is assessed by examining the relationship of confidence scores to (a) performance accuracy scores from the Reading and Listening tests, (b) personality traits, and (c) questionnaire measures of metacognitive awareness. In accordance with previous findings with different test batteries, we hypothesize that confidence scores from Reading and Listening tests will measure a psychological trait that is different from traits from these three distinct domains. Exploratory factor analysis is used to check if a separate confidence factor, independent of accuracy and metacognitive factors, can be identified. In addition, measures of Big Five personality traits are correlated with performance accuracy and confidence scores. If confidence is indeed a trait separate from personality traits, it can be expected that correlation between confidence and personality traits would not be too high and perhaps comparable to the correlation of accuracy scores with personality traits.

*Predictive validity* of confidence measures is examined using regression analyses and correlations. Several performance accu-

racy scores are used as criterion measures: scores from TOEFL iBT Writing and Speaking sections, TOEFL iBT total scores, standardized measures of academic performance (SAT and ACT), HS-GPA, and two additional cognitive tests (numeracy and overclaiming). Although it is hypothesized that both performance accuracy and confidence scores from Reading and Listening tests are good predictors of these criteria, it is expected that confidence scores have incremental validity over and above accuracy scores for at least some criterion measures. Confidence scores are also related to another criterion measure: the amount of change in performance accuracy between two occasions of testing. It is hypothesized that people who have low confidence will change their answers between two occasions of testing to a greater extent than those with high confidence. This hypothesis implies negative correlation between confidence scores and absolute values of change scores between two occasions of testing.

Second, we explore group differences (i.e., gender and ethnic differences) on accuracy, confidence, and bias scores from Reading and Listening tests. The presence of these differences can be interpreted as additional evidence for construct validity of confidence and bias scores. In accordance with previous findings, it is expected that men show greater overconfidence bias than women. In other words, it is expected that women are better calibrated than men. With respect to hitherto unexplored ethnic differences between Whites, Blacks, and Hispanics, no definite prediction can be made. However, indirect evidence can be used to hypothesize that groups with low performance accuracy scores show higher overconfidence bias than those with high accuracy scores. Thus, ethnic groups that show lower performance accuracy are expected to show greater overconfidence bias.

## METHOD

### Participants

As mentioned above, the data for the present study were collected as a part of the validation process for TOEFL iBT. Participants were recruited for the validation study and were paid \$100 for taking part in this research. This was not a high-stakes assessment for them. To save time and effort and collect data on additional noncognitive variables, including measures of self-confidence, we decided to use the same sample of participants.

Participants ( $N = 824$ ) were recruited from two types of colleges. One group ( $N = 371$ ) came from nine 2-year community colleges.<sup>3</sup> The other group ( $N = 453$ ) came from twelve 4-year colleges, from across the United States. All participants were native speakers of English. There were 304 male and 518 female

<sup>2</sup> This suggestion is supported by the present data. For the subsamples who reported their SAT and ACT scores, the correlations are:  $r_{\text{TOEFL,SAT}} = .54$ ,  $r_{\text{TOEFL,ACT}} = .65$ ,  $r_{\text{TOEFL,HSGPA}} = .33$  (see Table 10 for the  $N$ s).

<sup>3</sup> The total  $N$  for this study was, in fact, 950. Preliminary screening of the data showed that a proportion of participants produced patterns of responses indicating careless responding (e.g., the same answer was provided for all items in a given instrument). This tendency was particularly pronounced with the instruments administered in the afternoon session, not with TOEFL iBT scores. Our criterion for excluding participants was the presence of evidence for careless responding in 3 out of 28 instruments administered in the afternoon session, some of which are used in the present report.



participants. In terms of ethnic composition, the sample consisted of Whites ( $N = 605$ ), African Americans ( $N = 112$ ), Hispanics ( $N = 60$ ), and Others ( $N = 46$ ). The mean age of the sample was 19.6 years ( $SD = 3.2$ ).

### Procedure

Participants were administered the full TOEFL iBT exam in the morning and were asked to attend another session in the afternoon on the same day. The order of TOEFL iBT subtests in the morning session was: Reading, Listening, Writing, Speaking. The data from this session were intended for validation purposes.

The afternoon session started with the repeated (untimed) test consisting of the selected Reading and Listening items with confidence ratings attached to each item (see the *Instruments* section of this article). The time interval between beginning of morning and afternoon presentation of the same tests was 4 hr, shorter than what is usually done in studies of test-retest reliability. However, it should be kept in mind that there was intensive cognitive activity (i.e., taking of Writing and Speaking tests) between the two sessions, which may have affected memory traces for the repeated items. Thus, the increase in the overall performance accuracy in the second session is expected to be relatively small. The effect of prior exposure to the same items on the confidence scores is unknown at present. One would expect some increase in confidence and improvement in its reliability but the size of these effects is hard to estimate.

After completing the Reading and Listening tests, the participants took a battery of 28 instruments. These additional instruments were a measure of cognitive ability, personality, metacognition, interests, and emotional intelligence, as well as social attitudes, values, and social norms. Noncognitive variables require less mental effort than TOEFL iBT. With a lunch break between the morning and afternoon sessions and two breaks in the afternoon session, fatigue was not deemed to have a major influence on performance. Nevertheless, the effects of fatigue on test performance cannot be discounted. Depending on participants' speed of working through the paper-and-pencil noncognitive measures, the afternoon session lasted approximately 3 hr. In this article, we only report results on the accuracy and confidence scores obtained from the repeated items on the TOEFL iBT Reading and Listening sections. The validation process for confidence scores includes the measures of cognitive ability, personality, and metacognition selected from the entire battery of 28 measures.

### Instruments

#### TOEFL iBT

TOEFL iBT (Form B) consists of four sections: Reading, Listening, Speaking, and Writing. The description of these four sections follows.

**Reading.** The Reading section has three sets of items, each associated with a common reading passage of approximately 700 words. It contains a total of 40 items (12, 14, and 14 in Sets 1, 2, and 3, respectively). The examinees are allowed to spend 60 min completing the Reading section. Thirty-seven items are four-option multiple-choice items. One item in each set is an open-ended item. The work reported in this article is based on the

multiple-choice items in the first two item sets. Thus, Reading Version 1 consisted of 11 items from the first set and Reading Version 2 consisted of 13 items from the second set. Figure 1 presents a screen capture of an item from the Reading section and confidence ratings scale associated with the item.

**Listening.** The Listening section has 34 items in six item sets, with two item sets based on conversation and four item sets based on lectures on academic topics. Each stimulus is approximately 3 min long and is followed by 5 items with four multiple-choice options. A lecture stimulus lasts 3 to 5 min, followed by 6 items. After listening to the prompts, the participants can spend up to 20 min responding to all Listening items. Given a constraint on testing time in the afternoon testing session, we selected only conversation-based Listening items (17 items). Listening Version 1 contains 11 items from two different conversations, and Listening Version 2 contains 6 items from another conversation.

**Speaking.** The Speaking section consists of six items. Two items require examinees to express opinions on familiar topics. The other four items integrate speaking with other language tasks. Two of the four integrated tasks combine listening and speaking skills, which requires examinees to listen to a short dialogue and then to talk about the content of what they heard. The remaining two are Reading/Listening/Speaking tasks, which require examinees to read a short passage, listen to a dialogue that pertains to the passage, and then speak about what they have read and heard. For each of these items, examinees are given 15–30 s to prepare and 45–60 s to respond. Examinees' responses to each item are scored on a scale of 0–4 by trained raters. The raw Speaking score is a sum of the points across all six items; thus, examinees' raw Speaking scores range from 0 to 24.

**Writing.** The Writing section includes two items, one of which is an independent writing task, and the other is a writing task integrated with a reading task. The independent writing task requires examinees to argue for an opinion on a topic. The integrated writing task requires the examinees to read a text, listen to a lecture that pertains to the topic, and then write on a specific topic that is based on what examinees have read and heard. The testing time for the Writing section is 20 to 50 min for the independent writing task and 30 min for the integrated writing task. Examinees' responses to each item are scored on a scale of 0–5 by trained raters. The raw Writing score is a sum of the points earned on the two items; thus, the raw score ranges from 0 to 10.

The raw scores from each section of TOEFL iBT are converted to scaled scores of 0 to 30. The total TOEFL iBT score is a simple sum of the four scaled scores, ranging from 0 to 120.

#### Additional Ability Measures

Two ability measures, apart from TOEFL iBT, were used in this study:

**Numeracy test.** The National Adult Literacy Survey developed by ETS contains a numeracy test to assess examinees' ability to use and manipulate numerical information in a real-world context (Kirsch et al., 2001). Seventeen items from this test were used in the present study.

**Overclaiming test.** This 45-item instrument (see Paulhus & Harms, 2004) assesses respondents' overclaiming familiarity with historical names and events, social sciences, and physical sciences. Each item is on a scale ranging from 0 (*never heard of it*) to 6



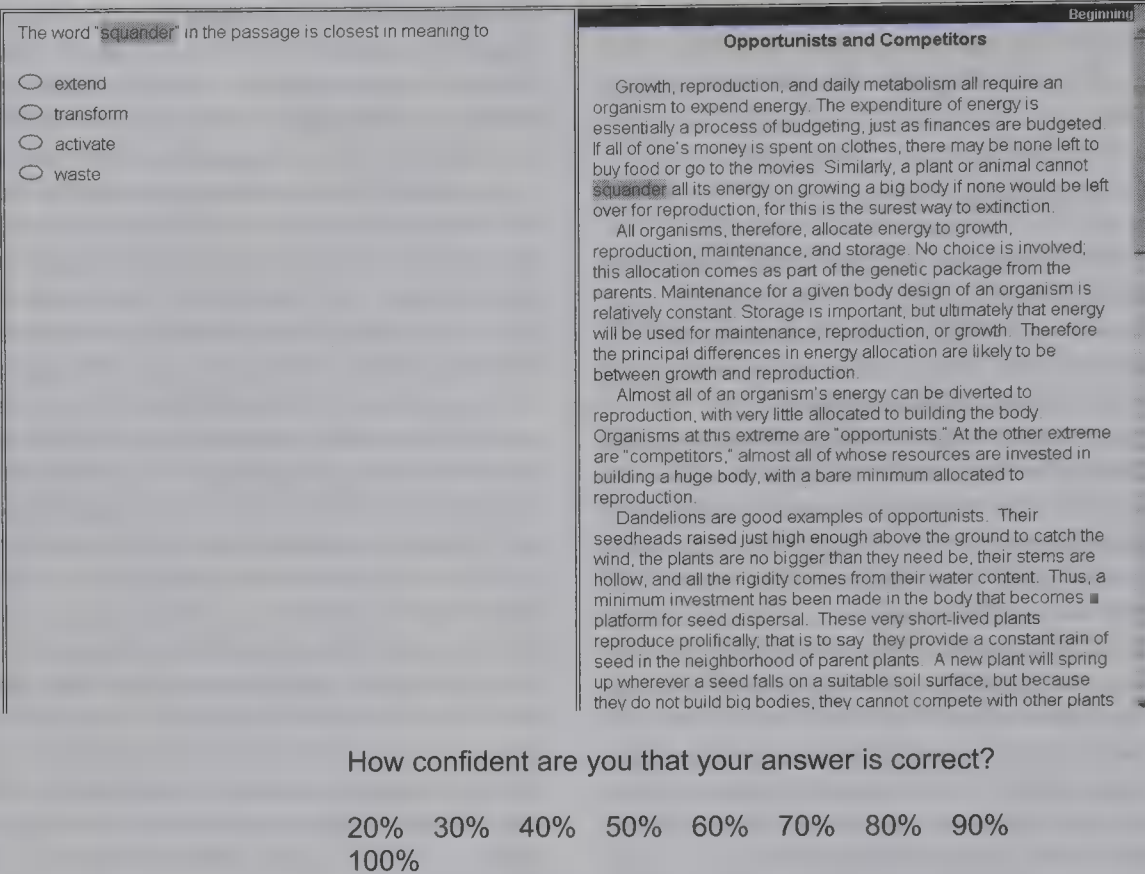


Figure 1. Screen capture of an item from the Test of English as a Foreign Language, Internet-Based Test, Form B. This item was used in a pilot study in 2006; it is not part of the test as currently administered. After providing the answer to an item, participants were asked to answer the confidence question.

(*know it very well*). Within each of three categories, 3 out of every 15 items are foils—that is, they do not actually exist. Hence, any degree of familiarity with them constitutes overclaiming. The data on overclaiming are typically analyzed by the signal detection theory (Paulhus & Harms, 2004). Signal detection analysis exploits all of the data in the calculation of separate indexes for *accuracy* and *response bias*. The best known formula for accuracy is *d'*. It assumes that an accurate individual is not the one scoring the most hits, but the one showing the best discrimination in claiming to recognize real items relative to foils. We did not employ response bias measures in this study.

Metacognitive Inventories

Three questionnaires were employed to assess metacognition.

*Metacognitive Awareness Inventory (MAI).* Ten items were chosen from the original 52 items of the MAI (Schraw & Dennison, 1994). These items showed satisfactory psychometric properties in a pilot for the present study. In the pilot study, all 52 items of MAI defined a single (general) metacognitive factor. The 10 selected items had the highest loadings on this general factor. Cronbach's alpha in this study is .76. The MAI contains questions on students' self-perceptions as learners, their strong and weak points, and their learning strategies and conditions under which they can learn most effectively.

*Memory and Reasoning Competence Inventory (MARCI).* This is a measure of self-concepts of memory and reasoning (Kleitman, 2003; Kleitman & Stankov, 2007). The inventory con-

sists of 16 items, 8 items for each component. The items assess people's perception of their reasoning and memory abilities relative either to other cognitive abilities they possess or to other people's cognitive abilities. The process of scale development is based on a model of self-concept items (Marsh & Shavelson, 1985; see also Marsh, 1986). The respondents evaluate the extent to which each statement describes them using a 6-point Likert scale ranging on a 6-point gradation from *false* (scaled at 1) to *true* (scaled at 6). Reasoning and memory items are intermixed, and separate scores for these two components are calculated. Kleitman (2003) has shown that 8 items of memory and 8 items of reasoning define two separate factors. Examples of memory items include "Compared to other intellectual abilities (i.e., attention, reasoning), my memory is good," and "My memory is above average." Examples of reasoning items include "I feel confident when solving problems that require reasoning skills," and "I can reason better than the average person." In the present study, Cronbach's alpha for the Reasoning component of MARCI is .73; for the Memory component, it is .79.

Personality Measures

For a personality measure, we used the Big Five Personality Inventory scales for Extraversion, Conscientiousness, Agreeableness, Emotional Stability, and Openness with reliability coefficients in this study of .81, .75, .75, .81, and .78 respectively. These scales are available from the International Personality Item Pool (n.d.).

Table 1  
*Arithmetic Means for Accuracy, Confidence, and Bias Scores for TOEFL iBT Reading and Listening Sections*

Test	<i>N</i>	Accuracy		Confidence <i>M</i>	Bias <i>M</i>
		<i>M</i> ( <i>SD</i> )	% correct		
1st presentation: TOEFL iBT					
Reading 1	11	8.55 (2.12)	78.78	—	—
Reading 2	13	9.38 (2.58)	72.15	—	—
Listening 1	11	8.84 (2.06)	80.36	—	—
Listening 2	6	4.94 (1.17)	82.33	—	—
2nd presentation: TOEFL iBT with confidence scale					
Reading 1	11	8.78 (2.19)	79.82	88.47	8.65
Reading 2	13	9.44 (2.56)	72.54	87.37	14.83
Listening 1	11	8.98 (2.05)	81.64	87.21	5.57
Listening 2	6	5.01 (1.19)	85.00	90.48	5.48

*Note.* *N* = 824. Dashes indicate that information was not available. TOEFL = Test of English as a Foreign Language; iBT = Internet-Based Test.

### Outcome Measures

Participants were also asked to provide their scores on standardized tests (either the SAT or ACT) and their HS-GPA. These are self-reported scores that have not been checked for accuracy.

## RESULTS

### Descriptive Statistics on the Accuracy and Confidence Scores

Descriptive statistics for accuracy and confidence scores are provided in Table 1. The information about the TOEFL iBT first presentation in Table 1 was obtained from the selected items during the full operational TOEFL iBT testing session in the morning. The information about the second presentation was obtained from the same items given in the afternoon session, which included participant-selected confidence ratings for each item. Raw scores were transformed into percentage-correct scores by dividing the raw score by the number of items in the test and multiplying by 100. The means column for confidence in Table 1 shows the average confidence score across the items in each scale. The bias column in Table 1 presents the difference between the means for confidence and the percentage-correct scores.

Four observations are noteworthy in Table 1. First, there are remarkably small differences in the mean performance between the morning and afternoon sessions. What is surprising is that the absence of time restriction in the afternoon session did not affect performance to any significant degree, at least at the group-mean level. Second, the Reading section was more difficult than the Listening, although both were somewhat easy as expected from the sample of native English speakers—the average difficulty levels range between 72% and 85%. The effect of a test's easiness on confidence scores is to reduce the number of low confidence ratings in the test. However, the work of Stankov and Crawford (1996, 1997) and Kleitman and Stankov (2001, 2007) suggests that test difficulty does not affect the nature of the relationship between accuracy and confidence. Third, the mean variations across four sets of tasks (i.e., Reading 1 and 2 and Listening 1 and 2) were

more salient in the accuracy scores than the confidence scores. In other words, the participants' confidence scores were more constant throughout the different tasks. Fourth, Reading 2 shows an unusually high bias score in the present data.<sup>4</sup> Previous studies with cognitive tests indicated that bias scores greater than 10% could be treated as nonnegligible (Kleitman & Stankov, 2001; Stankov & Crawford, 1996, 1997).

### Reliabilities of the Accuracy and Confidence Scores

Three types of reliability coefficients are presented in Table 2: (a) Cronbach's alphas for the accuracy and confidence scores, (b) parallel form (i.e., correlations between the accuracy scores on Versions 1 and 2 for the Reading and Listening sections), and (c) test-retest reliabilities (i.e., correlations between the accuracy scores from the two administrations).

The reliability coefficients for both the accuracy and the confidence scores are all at an acceptable level for research purposes, except for Listening 2, which contains only six items. It is noteworthy that reliabilities of the confidence scores are consistently high across different tasks and different versions. The confidence scores show higher reliabilities than the accuracy scores.

### Effect of Confidence on Changes in Accuracy Scores From Test to Retest

This section explores the cause of less than perfect correlations in the accuracy scores between the two testing sessions as a function of participants' confidence levels. We obtained absolute change scores from the differences in the accuracy scores between the morning and afternoon sessions. The absolute change scores indicate the amount of change irrespective of whether there was a

<sup>4</sup> One of the items in the Reading Version 2 was based on a partial scoring method—i.e., instead of 0, 1 scoring, a 0, 1, 2, 4 scoring key was employed. When we removed this item from the analyses, the overall bias score for Reading Version 2 was reduced by 2.94 percentage points, closer to the borderline value for noteworthy bias scores.



Table 2  
Reliabilities for TOEFL iBT Reading and Listening Scores With and Without Confidence Scale

Test	N	$\alpha$	Accuracy		Confidence $\alpha$
			Parallel	Test-retest	
1st presentation: TOEFL iBT					
Reading 1	11	.73	—	—	—
Reading 2	13	.79	.70	—	—
Listening 1	11	.75	—	—	—
Listening 2	6	.62	.55	—	—
2nd presentation: TOEFL iBT with confidence scale					
Reading 1	11	.82	—	.85	.91
Reading 2	13	.79	.72	.78	.94
Listening 1	11	.78	—	.71	.94
Listening 2	6	.72	.59	.68	.90

Note. Dashes indicate that information was not available. TOEFL = Test of English as a Foreign Language; iBT = Internet-Based Test.

reduction or increase in accuracy scores between these two sessions. The arithmetic means for the absolute change scores are shown in Table 3. Comparison between the means for the absolute change scores shows that changes occurred more frequently on the Reading section, which is more difficult, than on the Listening section. On the other hand, the correlation between Listening and Reading absolute change scores is somewhat high ( $r = .60$ ), indicating that the tendency for participants' scores to change between the two testing sessions is consistent across the two sections. The correlations between the absolute change scores and the confidence scores are negative, with the correlations in the upper .20s—more confident people tend to have lower absolute change scores. Thus, confidence is possibly one source of unreliability in the accuracy test scores, reflected in test-retest correlations. This may be because low confidence and guessing go hand in hand, and guessing, of course, lowers a test's reliability. Another interpretation of this finding is in terms of predictive validity—confidence scores may be predictors of repeated taking of high-stakes tests. People who show low confidence may be more prepared than those with high confidence to take a test over and over in the hope of improving their scores.

Table 3  
Pearson Product-Moment Correlations Between Absolute Change Scores Between Test and Retest and Confidence Scores on TOEFL iBT Reading and Listening Sections

Variable	Absolute change scores		Confidence scores	
	Reading	Listening	Reading	Listening
Absolute change scores				
Reading	—			
Listening	.60	—		
Confidence scores				
Reading	-.29	-.27	—	
Listening	-.18	-.26	.77	—
M (SD)	2.22 (3.02)	1.58 (2.20)		

Note. TOEFL = Test of English as a Foreign Language; iBT = Internet-Based Test.

### Factor Analyses of the Confidence Scores

This section explores whether the confidence factor stands separate from cognitive abilities and metacognition. Two analyses were carried out. The first factor analysis employed four sections of the TOEFL iBT exam that were administered in the afternoon session. This analysis was based on eight variables—accuracy and confidence scores from four measures (Listening 1 & 2 and Reading 1 & 2). The second analysis included the first battery of sections as well as other cognitive ability tests—numeracy and overclaiming  $d'$ —and the questionnaire measures of metacognition (MARCI measures of reasoning and memory, and MAI). The reason for carrying out two factor analyses is our assumption that four accuracy and four confidence scores derived from two tests that are purported to measure essentially the same construct (i.e., language proficiency) are conceptually at the same level as other cognitive ability tests employed in the second analysis. In other words, Versions 1 and 2 of the same test are parallel forms of estimates and their sum corresponds to a single, say, numeracy score.

#### First Factor Analysis

Table 4 presents correlations between the accuracy and confidence scores from the two versions of the TOEFL iBT Reading and Listening sections. All correlations are moderately high and positive. Several observations are noteworthy. First, the confidence scores correlate higher among themselves than do the accuracy scores. Second, for both accuracy and confidence scores, the correlations between the two sections (i.e., between Versions 1 and 2 of the same test) are higher than the correlations between Reading and Listening sections. Third, correlations between the accuracy and confidence scores from the same task range from .445 to .605. They are higher than the correlations between accuracy and confidence scores from the different versions of the same tasks (ranging from .358 to .574). Fourth, there are somewhat stronger associations between Reading accuracy and confidence scores (ranging from .469 to .605) than between Listening accuracy and confidence scores (ranging from .358 to .490).

Tables 5 and 6 provide a factor pattern matrix on the accuracy and confidence scores. The exploratory factor solution was ob-

Table 4  
Pearson Product-Moment Correlations Between Accuracy and Confidence Scores From TOEFL iBT

Variable	Accuracy scores				Confidence scores			
	Reading 1	Reading 2	Listening 1	Listening 2	Reading 1	Reading 2	Listening 1	Listening 2
Accuracy scores								
Reading 1	—							
Reading 2	.781	—						
Listening 1	.552	.595	—					
Listening 2	.569	.616	.684	—				
Confidence scores								
Reading 1	.605	.574	.431	.499	—			
Reading 2	.469	.519	.366	.450	.824	—		
Listening 1	.350	.360	.445	.490	.704	.768	—	
Listening 2	.282	.291	.358	.480	.629	.717	.822	—

Note. TOEFL = Test of English as a Foreign Language; iBT = Internet-Based Test.

tained by applying the root-one criterion and using the maximum likelihood estimation procedure followed by promax rotation. It clearly shows that two factors—verbal comprehension (i.e., accuracy scores from the TOEFL iBT Reading and Listening sections) and confidence—exist in these data. These two factors account for 77.44% of total variance of the eight-test battery. Factor intercorrelation (.578) is within expectations because typical correlations between confidence and performance factors range between .40 and .60.

### Second Factor Analysis

Table 7 shows correlations between the accuracy scores on cognitive ability tests (TOEFL iBT Reading and Listening sections, Versions 1 and 2 combined, the numeracy and overclaiming  $d'$  tests), the metacognitive measures (memory inventory, reasoning inventory, and metacognitive awareness inventory), and the confidence scores from TOEFL iBT Reading and Listening sections, Versions 1 and 2 combined. All the correlations in Table 7 are positive and moderate in size (except for essentially zero correlation between metacognitive awareness and overclaiming  $d'$

scores). As expected, the correlations in Table 7 are lower overall than those presented in Table 4.

Tables 8 and 9 show the factor pattern matrix and the factor correlation matrix based on the correlation matrix presented in Table 7. We report maximum likelihood solution followed by promax rotation. Using the root-one criterion, three factors were extracted. These factors account for 65% of the total variance. They are:

*Acculturated knowledge (Gc).* The accuracy scores from all four tests of cognitive abilities load on this factor. Although three of the four tests that load on this first factor involve verbal abilities, loading from the numeracy test indicates that the processes captured by this factor are somewhat broader than verbal comprehension, perhaps tapping crystallized abilities. Reading confidence also has a small loading on this factor.

*Confidence.* The confidence scores from the TOEFL iBT Reading and Listening sections load on this factor. Unfortunately, this is a doublet due to the absence of other measures of confidence in this study. However, previous work has shown that the confidence factor can be reliably extracted from a larger number of cognitive tasks (see Kleitman & Stankov, 2001; Pallier et al., 2002; Stankov, 1998, 1999, 2000; Stankov & Crawford, 1996, 1997).

*Metacognition.* All three questionnaire measures of metacognitive processes load on this factor.

The correlation between the acculturated knowledge ( $Gc$ ) and confidence factors is moderate in size ( $r = .552$ ), about the same as the one reported between verbal comprehension and confidence in Table 5. The metacognitive factor shows lower correlations with both  $Gc$  and confidence factors ( $r = .379$  and  $.317$ , respectively).

Table 5  
Exploratory Factor Analysis of the Correlations Between Accuracy and Confidence Scores

Variable	Factor	
	Confidence	Verbal comprehension
Accuracy scores		
Reading 1		.918
Reading 2		.967
Listening 1		.598
Listening 2	.215	.570
Confidence scores		
Reading 1	.593	.358
Reading 2	.753	
Listening 1	.962	
Listening 2	.958	

Note. Empty cells indicate that factor loading is lower than .200.

Table 6  
Factor Correlation Matrix

Factor	Confidence	Verbal comprehension
Confidence	—	
Verbal comprehension	.578	—



Table 7  
Pearson Product-Moment Correlations Among Accuracy and Confidence Scores and Metacognitive Inventories

Variable	Accuracy scores				Confidence scores		Metacognitive inventories		
	Reading 1 & 2	Listening 1 & 2	Numeracy	Overclaiming <i>d'</i>	Reading 1 & 2	Listening 1 & 2	Memory	Reasoning	Metacognitive awareness
	1	2	3	4	5	6	7	8	9
Accuracy scores									
Reading 1 & 2	—								
Listening 1 & 2	.693	—							
Numeracy	.616	.532	—						
Overclaiming <i>d'</i>	.375	.304	.355	—					
Confidence scores									
Reading 1 & 2	.594	.501	.437	.254	—				
Listening 1 & 2	.349	.495	.306	.168	.774	—			
Metacognitive inventories									
Memory	.375	.137	.157	.081	.196	.145	—		
Reasoning	.164	.268	.318	.169	.339	.254	.466	—	
Metacognitive awareness	.322	.019	.028	-.067	.102	.088	.271	.302	—

*Correlations Between TOEFL iBT Reading and Listening Confidence Scores and Accuracy Scores From TOEFL iBT, SAT, ACT, and HS-GPA*

Table 10 presents correlations between a composite confidence score (i.e., the sum of the confidence scores from the TOEFL iBT Listening and Reading sections) and other cognitive test scores (i.e., total TOEFL iBT, SAT, and ACT) and HS-GPA.

As expected, correlations between confidence scores and scores from the TOEFL iBT Reading (.499) and Listening (.539) sections are higher than correlations between confidence scores and scores from the TOEFL iBT Speaking (.338) and Writing (.414) sections. Confidence scores correlate moderately with total TOEFL iBT scores ( $r = .553$ ), which is comparable to the correlations between the confidence and accuracy factors in Tables 6 and 9. Less than half of participants provided self-reported SAT and ACT scores. For these subsamples of participants, the correlation of self-

reported SAT and ACT scores with confidence scores is rather low ( $r = .271$  and  $.238$ ). Self-reported HS-GPA shows the lowest correlation with confidence scores ( $r = .159$ ).

*Correlations Between Confidence Scores and Big Five Personality Traits*

Table 11 presents correlations between Big Five personality dimensions and total TOEFL iBT and composite confidence scores. The accuracy and confidence scores based on the TOEFL iBT exam show a moderate relationship with openness. This finding is in agreement with the findings of Pallier et al. (2002). We obtained somewhat higher correlations of agreeableness with both the accuracy and confidence scores than previous studies have reported (Kleitman & Stankov, 2007; Pallier et al., 2002). Generally speaking, the pattern of personality traits' correlations is similar for the accuracy and confidence scores. What is particularly noteworthy is that personality traits show slightly higher average correlations with the accuracy scores than with the confidence scores. In previous studies, the construct of confidence has been linked to personality traits more often than any other constructs (see Blais, Thompson, & Baranski, 2005; Klayman et al., 1999; Kleitman & Stankov, 2001; Pallier et al., 2002). Our data in this study suggest that personality traits have the same relationship to cognitive abilities as they do to confidence, or slightly closer.

Summary of Structural Findings

We have shown that confidence can be measured reliably. Our findings suggest that although moderately correlated with cogni-

Table 8  
Exploratory Factor Analysis of the Correlations Between Accuracy and Confidence Scores and Metacognitive Inventories

Variable	Factor		
	Acculturated knowledge	Confidence	Metacognition
Accuracy scores			
Reading 1 & 2	.962		
Listening 1 & 2	.640		
Numeracy scores	.630		
Overclaiming <i>d'</i> scores	.409		
Confidence scores			
Reading 1 & 2	.314	.618	
Listening 1 & 2		.994	
Inventories			
Memory			.626
Reasoning			.724
Metacognitive awareness			.453

Note. Empty cells indicate that factor loading is lower than .300.

Table 9  
Factor Correlation Matrix

Factor	Acculturated knowledge	Confidence	Metacognition
Acculturated knowledge	—		
Confidence	.552	—	
Metacognition	.379	.317	—

tive abilities, metacognition, and personality traits, a confidence factor is distinct and cannot be reduced to either ability or personality domains. There is nothing in the present data that challenges the conclusion that confidence is indeed a separate trait.

### Group Differences in Confidence: Gender, Ethnicity, and College Type

Table 12 displays the mean levels of the accuracy, confidence, and bias scores for the TOEFL iBT Reading and Listening sections for the total number of participants and male and female participants. Positive signs for mean bias scores indicate overconfidence at the group level of analysis. Men tend to show greater overconfidence bias than women. In this sample, the magnitude of the overconfidence bias for both men and women is somewhat lower than what was reported by Pallier (2003). Thus, Cohen's *d*'s for the bias scores presented in Table 12 are 0.15 for Reading and 0.25 for Listening.

Table 13 presents ethnic group means (Whites, African Americans, and Hispanics) for the accuracy, confidence, and bias scores. The Hispanic group is in the middle on all three scores. It is evident that bias is greater for African American groups than it is for either White or Hispanic groups. On both the TOEFL iBT Reading and Listening sections, Whites show the smallest bias scores. In the next section, we elaborate on the finding that *F* tests for confidence are smaller than *F* tests for accuracy. Cohen's *d*'s for the Reading bias scores presented in Table 13 are 0.83 (Whites vs. African American), 0.46 (Whites vs. Hispanics), and 0.37 (African Americans vs. Hispanics). Listening bias scores produced similar effect sizes.

We broke down the African American sample further by college type—those who attend 2- or 4-year colleges. Table 14 shows the means for the composite accuracy, confidence, and bias scores for these subgroups. The differences on accuracy scores are not significant, whereas the differences on confidence are significant. African Americans in 2-year colleges show stronger bias than those attending 4-year colleges,  $t(110) = 4.07, p < .01$ . For Whites and Hispanics, the difference between 2- and 4-year colleges on bias was not statistically significant.

Table 10  
*Pearson Product-Moment Correlations Between TOEFL iBT Reading and Listening Confidence Scores and Various Accuracy Scores*

Variable	TOEFL iBT Reading and Listening confidence score
Accuracy scores	
Reading 1 and 2	.499
Listening 1 and 2	.539
Speaking	.338
Writing	.414
TOEFL total score	.552
Self-reported SAT <sup>a</sup>	.271
Self-reported ACT <sup>b</sup>	.348
HS-GPA	.159

Note. TOEFL = Test of English as a Foreign Language; iBT = Internet-Based Test; HS-GPA = high school grade point average.

<sup>a</sup> Subsample of  $N = 384$ . <sup>b</sup> Subsample of  $N = 342$ .

Table 11  
*Pearson Product-Moment Correlations Between Big Five Factors and Reading and Listening Confidence Scores and TOEFL iBT Total Score*

Big Five factors	Accuracy scores	Confidence scores
	TOEFL iBT total score	Reading and Listening sections
Extraversion	.037	.040
Agreeableness	.343	.234
Conscientiousness	.138	.158
Emotional stability	.046	.119
Openness	.390	.332
Average	.191	.177

Note. TOEFL = Test of English as a Foreign Language; iBT = Internet-Based Test.

### Hard–Easy Effect and Overconfidence Bias

Tables 12, 13, and 14 show a consistent pattern: The between-group variability is lower for the confidence scores than it is for the accuracy scores. For instance, Table 13 shows that Whites on the TOEFL iBT Listening section demonstrate the highest accuracy score of 85.27, whereas African Americans on the TOEFL iBT Reading section show the lowest accuracy score of 61.75. The range for accuracy scores between groups is 23.52. On the other hand, the highest confidence score on the TOEFL iBT Listening section, 89.52, is reported for Whites, whereas the lowest confidence score, 83.02, is reported for African Americans. The range for confidence scores is only 6.50. Thus, the main difference in the bias scores between these ethnic groups is due to greater differences in the accuracy scores, not due to the confidence scores. Although groups of people who receive low accuracy scores express lower confidence in their answers, group differences in confidence scores do not change to the same order of magnitude as accuracy scores. Thus, bias is attributed to people's inability to judge the degree of the failure to solve test problems. Confidence is somewhat resistant to change even when people perform poorly on a test.

Within the framework of the present study, it is not just group differences that point to the variability in test difficulty as the factor most responsible for changes in bias scores. Arithmetic means for the whole sample displayed in Table 1 indicate that the Reading section is more difficult than the Listening section, but the average confidence differs little between the Listening and Reading sections. As a result, the Reading section shows an overconfidence bias that is more than twice as large as in the Listening section. What is known as a *hard–easy effect* in decision-making literature (see Suantak, Bolger, & Ferrell, 1996) may be one possible explanation for these differences. Put simply, the magnitude of overconfidence bias depends on task difficulty, and overconfidence bias tends to be more pronounced as the task becomes more difficult.

Instead of focusing on item difficulty, as it is usually done in discussions of the hard–easy effect, we can focus on participants and point out again that the results presented in Tables 12–14 show that groups with lower ability have larger overconfidence bias. This is in accordance with the assumption that higher ability



Table 12  
Means for Accuracy, Confidence, and Bias Scores on TOEFL iBT Reading and Listening Sections, Versions 1 and 2, by Gender

Group	Reading 1 and 2			Listening 1 and 2		
	Accuracy	Confidence	Bias	Accuracy	Confidence	Bias
Total sample	76.18	87.92	11.74	82.54	88.84	6.30
Men <sup>a</sup>	74.57	87.58	13.01	79.88	88.27	8.39
Women <sup>b</sup>	77.23	88.16	10.93	84.19	89.20	5.01
<i>t</i> test <sup>c</sup>	2.040*	0.748	2.106*	4.204**	1.482	3.930**

Note. TOEFL = Test of English as a Foreign Language; iBT = Internet-Based Test.

<sup>a</sup> *n* = 304. <sup>b</sup> *n* = 518. <sup>c</sup> *df* = 821.

\* *p* < .05. \*\* *p* < .01.

people have increased experience and are better at encoding stimuli (Dougherty, 2001). However, as we shall argue in the last section of this article, this correlation between overconfidence bias and ability may be due to individual differences in the understanding of subjective probability concepts (Kleitman, 2003; Stankov, Lee, & Paek, in press).

### Predictive Validity of Confidence Scores

In this section, we examine evidence for incremental predictive validity of confidence scores. In all the analyses we report, the scaled accuracy scores for the TOEFL iBT Reading and Listening section were entered as the first block of predictors, and the overall confidence scores were added in the second block. The question of interest was whether confidence can account for additional variance in the criterion that remains after accuracy scores are entered into the regression equation.

Table 15 presents  $R^2$  statistics as a summary index for regression analyses. A table note in the last column in Table 15 indicates scores with significant differences in the  $R^2$  between the two regression models; these are known as *incremental  $R^2$* . The criterion measures are three types of cognitive outcomes: (a) scores derived from TOEFL iBT (the total TOEFL iBT score as well as Speaking and Writing section scores), (b) numeracy and overclaiming  $d'$  test scores, and (c) self-reported SAT scores and HS-GPA. Table 15 shows that the confidence scores derived from TOEFL iBT Reading and Listening sections contribute significantly to the prediction of the total TOEFL iBT score, as well as TOEFL iBT Speaking and Writing section scores. The confidence scores provide statistically significant improvement in predicting Writing and Speaking scores that were not used in the assessment of confidence.<sup>5</sup> Of the two tests administered independently of TOEFL iBT, the numeracy test shows incremental validity for the confidence scores. The correlations between the overclaiming  $d'$  test and the confidence scores were low to begin with (see Table 7:  $r = .254$  for the Reading confidence score and  $r = .168$  for the Listening confidence score). Thus, the absence of any evidence for incremental validity for overclaiming  $d'$  is not entirely surprising. No evidence for incremental validity is found for the two self-reported scores: SAT and HS-GPA. For all three criterion scores that did not show incremental validity, we do not have any information about their reliability. We may note, however, that they are much broader than accuracy and confidence measures obtained from Listening and Reading tests.

In summary, the confidence scores from the two sections of TOEFL iBT do provide incremental validity over and above their yoked accuracy scores to the three independent sets of cognitive tests scores—speaking, listening, and numeracy. We can conclude that people's confidence scores can predict their cognitive abilities in some measures even after controlling for the cognitive abilities that are used as the basis for measuring confidence level. Reported incremental predictive validity is important from the theoretical point of view, as very few noncognitive measures have shown any incremental validity over cognitive measures (Schmidt & Hunter, 1998) when training or job success are used as criteria. This is particularly important because both accuracy and confidence measures arise from the same cognitive act and, in the case of total TOEFL iBT, the amount already predicted by the Reading and Listening accuracy scores themselves is very high indeed (87.5%). However, because the amount of variance accounted for by the confidence scores in this study is one percent or less, the practical importance of this finding is minimal.

On the other hand, it is reasonable to assume that confidence may have a much more prominent role in predicting criteria other than those closely linked to cognitive performance. For example, within the educational setting, confidence may be a good predictor of dropout rates, much better than many cognitive measures in existence today (see Kuncel & Hezlett, 2007). The evidence presented in this article also suggests that confidence may be a good predictor of repeated test taking in high-stakes assessments. In business settings where decision-making is crucial, the role of confidence is likely to be even more important.

### DISCUSSION

In this section we summarize the main findings of this study by focusing on three issues: (a) structural findings and their implications for the understanding of miscalibration, (b) group differences and problems associated with the use of bias scores, and (c) validity issues in the measurement of confidence and possible use of confidence ratings in assessment. In the last section we mention questions that may lead to future research.

<sup>5</sup> In the regression analyses of this article, we also entered an interaction score (i.e., a multiple of accuracy and confidence scores). This interaction score did not contribute significantly to any of the criterion measures (rows) in Table 15.

Table 13

*Means for Accuracy, Confidence, and Bias Scores on TOEFL iBT Reading and Listening Sections, Versions 1 and 2, by Ethnicity*

Group	Reading 1 and 2			Listening 1 and 2		
	Accuracy	Confidence	Bias	Accuracy	Confidence	Bias
White <sup>a</sup>	79.32	89.02	9.69	85.27	89.52	3.92
African American <sup>b</sup>	61.75	83.02	21.27	68.87	84.98	16.11
Hispanic <sup>c</sup>	70.41	86.51	16.10	79.19	89.31	10.12
<i>F</i> test <sup>d</sup>	59.41**	17.28**	21.673**	51.24**	9.30**	19.41**

Note. TOEFL = Test of English as a Foreign Language; iBT = Internet-Based Test.

<sup>a</sup> *n* = 605. <sup>b</sup> *n* = 113. <sup>c</sup> *n* = 60. <sup>d</sup> *df* = 2769.

\*\* *p* < .01.

### Structural Findings and Their Implications

Our results are in agreement with previous work showing that confidence is an important psychological trait on the borderline between personality and cognitive abilities (Kleitman & Stankov, 2001, 2007; Pallier et al., 2002; Stankov, 1998, 1999, 2000; Stankov & Crawford, 1996, 1997). First, our results show that confidence measures have higher reliabilities than ability scores. Second, confidence should be considered as a separate trait, distinct from other traits such as ability and personality traits. The present study employed confidence scores based only on TOEFL iBT Reading and Listening scores, but previous work has shown that the confidence factor can be reliably extracted from a large number of cognitive tests. Our study also shows that confidence tends to be more closely related to cognitive abilities than it is to personality traits. This suggests that different types of confidence may exist—cognitive confidence, which is captured by the procedures of this article, versus social confidence, which can be measured as a part of personality. Third, although confidence appears to be conceptually related to processes captured by the questionnaire measures of metacognition, our evidence shows that confidence measures define a different factor. These findings provide support for construct validity of the confidence trait.

We believe that developmental evidence is needed to explain the existence of such a pervasive, i.e., general, cognitive confidence factor. Several studies are currently under way to examine the age at which children can be tested with the procedures described in this article. Our working hypothesis is that some factors within the school and family operate during the middle and high school periods to bring about this factor. The mechanisms are likely to be

Table 14

*Means for Accuracy, Confidence, and Bias Scores on Combined Reading 1 and 2 and Listening 1 and 2 Among African Americans by College Type*

College type	Accuracy	Confidence	Bias
2-year college	64.25	88.15	23.90
4-year college	66.39	81.15	14.76
<i>t</i> test <sup>a</sup>	1.53 ( <i>ns</i> )	2.97*	4.07*

<sup>a</sup> *df* = 110.

\* *p* < .05.

similar to those that lead to the appearance of the broad academic self-concept (see Guay, Marsh, and Boivin, 2003).

Identification of a new psychological trait is important from the individual differences point of view. Our findings, however, also have bearings on the explanation of miscalibration. Put simply, for a test made of items of the same difficulty, there are always individual differences in confidence that reflect a stable trait and therefore can be predicted using different measures of the same trait. For such a test confidence in addition to random error (Soll, 1996) can predict the amount of overconfidence bias.

### Group Differences and Problems With Measures of Overconfidence Bias

Three kinds of group differences have been identified: gender, ethnicity, and age. In this study, we found that men exhibit a stronger overconfidence bias than women. We also found differential overconfidence bias among ethnic groups, with African

Table 15

*R<sup>2</sup> Coefficients Showing Incremental Validity of Reading and Listening Confidence Scores in Predicting Various Accuracy Score Criteria Above and Beyond Reading and Listening Accuracy Scores*

Criteria	<i>R<sup>2</sup> from regression analysis</i>	
	Regression Model 1 predictors <sup>a</sup>	Regression Model 2 predictors <sup>b</sup>
TOEFL		
Total	.875	.877* <sup>c</sup>
Writing	.385	.395** <sup>c</sup>
Speaking	.269	.273* <sup>c</sup>
Numeracy	.401	.404* <sup>c</sup>
Overclaiming	.144	.145
SAT <sup>d</sup>	.307	.307
HS-GPA	.079	.079

Note. TOEFL = Test of English as a Foreign Language; iBT = Internet-Based Test; HS-GPA = high school grade point average.

<sup>a</sup> Reading and Listening accuracy scores. <sup>b</sup> Reading and Listening accuracy and confidence scores. <sup>c</sup> Statistically significant incremental validity change from the first model to the second model. <sup>d</sup> Subsample of *N* = 384.

\* *p* < .05. \*\* *p* < .01.



American students, particularly those attending 2-year colleges, showing the most pronounced overconfidence bias. Although one needs to be cautious in interpreting this finding because of the small sample, African Americans appear to be less aware of their level of performance on tests of verbal abilities than any other group in the present study, exhibiting the strongest bias. It seems plausible that overconfidence may lead to suboptimal effort and therefore contribute to the achievement gap. This may be an interesting avenue to pursue in future research on the topic. Finally, previous studies show age-related group differences: Older people are somewhat more confident than younger people (see Crawford & Stankov, 1996).

Our data also show that gender and ethnic group differences are smaller on confidence than they are on performance accuracy, implying that there is indeed a correlation between how difficult the task is for a given group and the size of the average overconfidence bias score. In other words, groups that have lower performance accuracy scores show higher overconfidence. This is in agreement with Dougherty's (2001) explanation of overconfidence bias in terms of increased experience and better encoding — i.e., the group with higher cognitive ability should show superior performance on tasks that depend on these processes.

However, the correlation between test difficulty (or participants' level of ability) and the amount of overconfidence bias needs to be examined carefully. The problem arises because of the issues associated with the measurement of subjective probability (confidence). For example, Stankov et al. (in press) point out that confidence ratings have a ceiling (100% confidence), and at least some of the subjects seem to interpret 50% confidence as chance performance, whereas others tend to use the full 20% to 100% range. Also, Kleitman (2003) argued that bias scores are indicative of systematic irrationality and correlate with several other biases identified by the heuristics and biases approach. Although these problems do not seem to seriously affect the measurement of confidence ratings per se, they do raise questions when we try to compare objective and subjective probability estimates. In addition, because bias scores are difference scores, they tend to have lower reliability than performance accuracy and confidence scores. It is for these reasons that the focus of this article is on confidence and not on measures of overconfidence bias.

### Validity of Confidence Scores

There is evidence for predictive validity of confidence scores. Thus, our results indicate that confidence ratings predict the amount of change in performance accuracy scores over two sessions of testing. This suggests that they may be indicative of repeated taking of high-stakes tests. Recent evidence also suggests that confidence scores are valuable in predicting some maladaptive personality styles, such as the feeling of being an impostor, also known as the "impostor phenomenon," which is characterized by a sense of inferiority, self-criticism, and a pervasive fear of the inability to replicate one's own success, despite previous evidence to the contrary. Want and Kleitman (2006) showed that people with a strong impostor disposition have lower confidence but not lower accuracy scores than those without such dispositions.

Our results also show an incremental validity of confidence scores for predicting cognitive performance on tests of writing, speaking, and numeracy. In particular, the incremental validity for

the numeracy test in this study underscores that confidence can be seen as a broad, task-independent trait. The absence of incremental validity evidence for self-reported SAT scores and HS-GPA may be due to the accuracy on self-reported test scores. Reliability of these self-reported scores is unknown. Possible low reliability of these self-reports may account for the low correlation between the confidence scores and both SAT scores and HS-GPA, and therefore be related to low incremental validity. From the practical point of view, the value of the obtained incremental validity is minimal.

The evidence for an incremental increase in the validity of confidence ratings suggests possible use of confidence if one wants to incorporate a noncognitive measure in service of selection, guidance, and intervention. It is the aspect of measurement employed in this study that makes confidence unique in comparison to other noncognitive measures. Of particular importance is the yoked nature of the confidence expressed in the answers to a specific test problem. This makes it difficult to fake an answer or to coach an examinee to answer in a particular way, which can occur in most other noncognitive measures. The presence of excessive bias scores can serve as a warning to the examiner that something may be amiss. The suggestion is that confidence ratings and bias scores may be useful as predictors of criteria other than cognitive performance. For example, they may be useful predictors of dropout rates or times to completion in graduate schools.

Finally, it is necessary to mention limitations on the findings reported in this article. These include possible effects of length of testing and fatigue on performance, the somewhat narrow range of ability levels of participants, and disproportionate cell sizes in ethnic group comparisons.

### Concluding Remarks

There are still quite a few unanswered questions about confidence. Some issues are relevant to the causal chain mentioned in relationship to achievement gap. For example, can general feedback on test performance reduce excessive confidence? Given that confidence has an incremental validity for predicting scores on the TOEFL iBT Speaking and Writing sections, does confidence have particular value in language learning? What is the relationship between confidence and personality measures other than the Big Five? For instance, how is ease in social interaction such as readiness to engage in public speaking related to cognitive confidence?

For over a century, the study of individual differences has focused on uncovering the dimensions that can be used to describe and perhaps understand the psychological make-up of the human species. Thus far, personality and ability domains have been mapped out reasonably well (see Carroll, 1993; Saucier & Goldberg, 2002). Every candidate for a new dimension in individual differences needs to be compared to what is known so far. At issue is a robust proof of its convergent and discriminant validity. Together with the findings from previous studies, the findings reported in this article indicate that confidence is indeed a psychological trait that is related to, but distinct from both personality and ability traits. Within the structure of all individual differences dimensions, confidence should be located between these two domains.



## References

- Angelis, J. C., Swinton, S. S., & Cowell, W. R. (1979). *The performance of non-native speakers of English on TOEFL and verbal aptitude tests* (TOEFL Report No. RR-03). Princeton, NJ: Educational Testing Service.
- Angoff, W. H., & Sharon, A. T. (1971). A comparison of scores earned on the Test of English as a Foreign Language by Native American college students and foreign applicants to U.S. colleges. *TESOL Quarterly*, 5, 129–136.
- Baranski, J. V., & Petrusic, W. M. (1999). Realism of confidence in sensory discrimination. *Perception and Psychophysics*, 61, 1369–1383.
- Baron, J. (2000). *Thinking and deciding* (3rd ed.). Cambridge, England: Cambridge University Press.
- Blais, A. R., Thompson, M. M., & Baranski, J. V. (2005). Individual differences in decision processing and confidence judgments in comparative tasks: The role of cognitive styles. *Personality and Individual Differences*, 38, 1701–1713.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York: Cambridge University Press.
- Crawford, J., & Stankov, L. (1996). Age differences in the realism of confidence judgments: A calibration study using tests of fluid and crystallized intelligence. *Learning and Individual Differences*, 6, 84–103.
- Dougherty, M. R. (2001). Integration of the ecological and error models of overconfidence using a multiple-trace memory model. *Journal of Experimental Psychology: General*, 130, 579–599.
- Festinger, L. (1943a). Studies in decision: I. Decision-time, relative frequency of judgment and subjective confidence as related to physical stimulus difference. *Journal of Experimental Psychology*, 32, 291–306.
- Festinger, L. (1943b). Studies in decision: II. An empirical test of a quantitative theory of decision. *Journal of Experimental Psychology*, 32, 411–432.
- Fullerton, G. S., & Cattell, J. M. (1892). *On the perception of small differences* (University of Pennsylvania Philosophy Series No. 2). Philadelphia: University of Pennsylvania Press.
- Gigerenzer, G., Hoffrage, U., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, 98, 506–528.
- Guay, F., Marsh, H. W., & Boivin, M. (2003). Academic self-concept and academic achievement: Developmental perspectives on their causal ordering. *Journal of Educational Psychology*, 95, 124–136.
- Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Sciences*, 1, 78–82.
- Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, 95, 528–551.
- International Personality Item Pool: A scientific collaboratory for the development of advanced measures of personality traits and other individual differences. (n.d.). Retrieved June 28, 2005, from <http://ipip.ori.org/>
- Johnson, D. C. (1977). The TOEFL and domestic students: Conclusively inappropriate. *TESOL Quarterly*, 11, 79–86.
- Juslin, P., & Olsson, H. (1997). Thurstonian and Brunswikian origins of uncertainty in judgment: A sampling model of confidence in sensory discrimination. *Psychological Review*, 10, 344–366.
- Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgments under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press.
- Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica*, 77, 217–273.
- Kirsch, I., Yamamoto, K., Norris, N., Rock, D., Jungleblut, A., & O'Reilly, P. (2001). *Technical report and data file user's manual for the 1992 National Adult Literacy Survey* (NCES-2001–457). Washington, DC: National Center for Education Statistics.
- Klayman, J., Soll, J. B., Gonzalez-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, 79, 216–247.
- Kleitman, S. (2003). *Self-confidence and self-monitoring aspects of metacognition, their nature and their place within rationality debates: An individual differences perspective*. Unpublished doctoral dissertation, University of Sydney, Sydney, Australia.
- Kleitman, S., & Stankov, L. (2001). Ecological and person-oriented aspects of metacognitive processes in test-taking. *Applied Cognitive Psychology*, 15, 321–341.
- Kleitman, S., & Stankov, L. (2007). Self-confidence and metacognitive processes. *Learning and Individual Differences*, 17, 161–173.
- Kuncel, N. R., & Hezlett, S. A. (2007, February 23). Standardized tests predict graduate students' success. *Science*, 315, 1080–1081.
- Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*, 20, 159–183.
- Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgments under uncertainty: Heuristics and biases* (pp. 306–334). Hillsdale, NJ: Erlbaum.
- Marsh, H. W. (1986). Verbal and math self-concepts: An internal/external frame of reference model. *American Educational Research Journal*, 23(1), 129–149.
- Marsh, H. W., & Shavelson, R. (1985). Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist*, 20(3), 107–123.
- Nelson, T. O., & Narens, L. (1992). Metamemory: A theoretical framework and new findings. In T. O. Nelson (Ed.), *Metacognition: Core readings* (pp. 117–170). Boston: Allyn & Bacon.
- Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25). Cambridge, MA: MIT Press.
- Olsson, H., & Winman, A. (1996). Underconfidence in sensory discrimination: The interaction between experimental setting and response strategies. *Perception and Psychophysics*, 58, 374–383.
- Pallier, G. (2003). Gender differences in the self-assessment of accuracy on cognitive tasks. *Sex Roles*, 48, 265–276.
- Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., et al. (2002). Individual differences in the realism of confidence judgments. *Journal of General Psychology*, 129, 257–300.
- Paulhus, D. L., & Harms, P. D. (2004). Measuring cognitive ability with the overclaiming technique. *Intelligence*, 32, 297–314.
- Saucier, G., & Goldberg, L. R. (2002). Assessing the Big Five: Applications of 10 psychometric criteria to the development of marker scales. In B. de Raad & M. Perugini (Eds.), *Big Five assessment* (pp. 29–58). Ashland, OH: Hogrefe & Huber.
- Sawaki, Y., Stricker, L., & Oranje, A. (2005). *Factor structure of the TOEFL Internet-based test (iBT)*. Unpublished manuscript.
- Schmidt, F., & Hunter, J. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262–274.
- Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, 19, 460–475.
- Soll, J. B. (1996). Determinants of overconfidence and miscalibration: The roles of random error and ecological structure. *Organizational Behavior and Human Decision Processes*, 65, 117–137.
- Stankov, L. (1998). Calibration curves, scatterplots and the distinction between general knowledge and perceptual tasks. *Learning and Individual Differences*, 10, 29–50.
- Stankov, L. (1999). Mining on the “no man's land” between intelligence and personality. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.), *Learning and individual differences: Process, trait, and content determinants* (pp. 315–337). Washington, DC: American Psychological Association.



- Stankov, L. (2000). Complexity, metacognition, and fluid intelligence. *Intelligence*, 28, 121-143.
- Stankov, L., & Crawford, J. D. (1996). Confidence judgments in studies of individual differences. *Personality and Individual Differences*, 21, 971-986.
- Stankov, L., & Crawford, J. D. (1997). Self-confidence and performance on tests of cognitive abilities. *Intelligence*, 25, 93-109.
- Stankov, L., & Kim, S. (2006). *Dimensions of cultural differences: Personality, social attitudes, values and social norms*. Unpublished manuscript.
- Stankov, L., Lee, J., & Paek, I. (in press). Realism of confidence judgments. *European Journal of Psychological Assessment*.
- Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. London: Erlbaum.
- Stricker, L. J. (2002). The performance of native speakers of English and ESL speakers on the computer-based TOEFL and GRE General Test. *Language Testing*, 21, 146-173.
- Suantak, L., Bolger, F., & Ferrell, W. R. (1996). The hard-easy effect in subjective probability calibration. *Organizational Behavior and Human Decision Processes*, 67, 201-221.
- Tobias, S., & Everson, H. T. (2000). Assessing metacognitive knowledge monitoring. In G. Schraw & J. C. Impara (Eds.), *Issues in the measurement of metacognition* (pp. 147-222). Lincoln, NE: Buros Institute of Mental Measurements.
- Tobias, S., & Everson, H. T. (2002). *Knowing what you know and what you don't: Further research on metacognitive knowledge monitoring*. (College Board Research Report No. 2002-3). Retrieved June 29, 2006, from <http://iume.tc.columbia.edu/downloads/tobias/CBR2001-3.pdf>
- Trow, W. C. (1923). The psychology of confidence. *Archives of Psychology*, 67, 47-71.
- Vickers, D. (1979). *Decision processes in visual perception*. New York: Academic Press.
- Want, J., & Kleitman, S. (2006). Feeling "phony": Adult achievement behaviour, parental rearing style and self-confidence. *Journal of Personality and Individual Differences*, 40(5), 961-971.

Received December 20, 2007

Revision received March 6, 2008

Accepted March 31, 2008 ■

# The Role of Passion for Teaching in Intrapersonal and Interpersonal Outcomes

Noémie Carbonneau and Robert J. Vallerand  
Université du Québec à Montréal

Claude Fernet  
Université du Québec à Trois-Rivières.

Frédéric Guay  
Université Laval

The purpose of this study was to determine the role of passion in teachers' burnout symptoms, work satisfaction, and perceptions of positive student classroom behaviors. The dualistic model of passion (Vallerand et al., 2003) proposes 2 types of passion: harmonious and obsessive. In previous studies, harmonious passion has been shown to lead to adaptive outcomes (e.g., well-being and satisfaction), whereas obsessive passion has been shown to lead to less adaptive outcomes (e.g., shame and negative affect). In this study, 494 teachers completed measures of passion for teaching and various outcomes associated with the teaching profession twice over a 3-month period. Results of a cross-lag model based on structural equation modeling revealed that increases in harmonious passion for teaching predicted increases in work satisfaction and decreases in burnout symptoms over time, while changes in obsessive passion were unrelated to such outcomes. In addition, increases in both harmonious and obsessive passion predicted increases in teacher-perceived adaptive student behaviors over time. Overall, the results of the present study suggest that passion for teaching is an important concept to consider in education.

**Keywords:** passion, teaching, student outcomes, burnout, work satisfaction

Passion is not a luxury, a frill, or a quality possessed by just a few teachers. It is essential to all good teaching. (Day, 2004, p. 11)

Teaching is a complex and demanding career that requires intense dedication. Daily difficulties teachers face include stress (Wilhelm, Dewhurst-Savellis, & Parker, 2000), inadequate support from the school administration, student discipline problems, and low salaries (Ingersoll, 2001). Every year, such harsh conditions make a number of teachers choose to leave the profession, which implies that those who stay committed to their position must feel a deep and genuine love for their job (Elliott & Crosswell, 2001). This is congruent with Day's (2004) claim that passion is essential in the teaching realm. Although the idea that passionate teachers have a positive impact on their students is widely accepted (Day, 2004; Fried, 1995; Patrick, Hisley, Kempler, & College, 2000), positions vary considerably as to the quality of personal outcomes

teachers derive from their passion. On one hand, some authors claim that passion in the workplace is the antidote to burnout because it characterizes people who are continually energized and reinvigorated by their work (Dlugos & Friedlander, 2001; Grosch & Olsen, 1994; Selder & Paustian, 1989). On the other hand, others consider that passion is one of the causes of burnout because it leads people to engage with too much intensity in a sphere of their life and neglect other life domains (Coulehan, 2002; Freudenberg, 1985; Marcil, 1991). Also, although the high level of commitment required by the teaching profession is a source of satisfaction for many teachers, others find the demands too great a burden (Day, 2004; Elliott & Crosswell, 2001) and report that teaching is "too absorbing" and "takes up too much of [their lives]" (Nias, 1989, p. 39).

How can passion lead a number of teachers to derive psychological well-being and satisfaction from their professional lives while driving others to be consumed by their teaching to the point where they end up suffering and experiencing negative emotions and burnout? Although much has been written about passion for teaching, very little empirical research and no formal theory or model seem to exist. The purpose of the present study was to attempt to address the above question by using a new conceptualization of passion (Vallerand et al., 2003).

## A Dualistic Model of Passion

Vallerand et al. (2003) defined passion as a strong inclination or desire toward an activity (e.g., one's job) that one likes (or even loves) and finds important and in which one invests time and energy. We further posit that the representation of an activity that

---

Noémie Carbonneau and Robert J. Vallerand, Laboratoire de Recherche sur le Comportement Social, Université du Québec à Montréal, Montreal, Quebec, Canada; Claude Fernet, Département des Sciences de la Gestion, Université du Québec à Trois-Rivières, Trois-Rivières, Quebec, Canada; Frédéric Guay, Chairholder of the Canada Research Chair on Motivation and Academic Success, Université Laval, Quebec, Canada.

The research was facilitated by grants from the Social Science and Humanities Research Council and by the Fonds de Recherche sur la Société et la Culture to Robert J. Vallerand.

Correspondence concerning this article should be addressed to Robert J. Vallerand, Laboratoire de Recherche sur le Comportement Social, Université du Québec à Montréal, Box 8888, Succursale Centre-Ville, Montréal, Québec, Canada H3C 3P8. E-mail: vallerand.robert\_J@uqam.ca



one likes and in which one engages on a regular basis will be incorporated in that person's identity to the extent that the activity is highly valued (Aron, Aron, & Smollan, 1992; Csikszentmihalyi, Rathunde, & Whalen, 1993), thereby leading to a passion for this activity. Such a passionate activity comes to be so self-defining that it represents a central feature of one's identity. For instance, those who have a passion for teaching do not merely say that they teach; they *are* teachers.

The passion model distinguishes two types of passion: harmonious and obsessive passion. There are two different processes by which an activity can be internalized in one's identity, each of them resulting in a specific type of passion. Harmonious passion emanates from an autonomous internalization (Deci & Ryan, 2000; Vallerand, 1997) of the activity into one's identity. Such an internalization occurs when an individual freely accepts an activity as important to him or her. No contingencies are attached to the passionate activity such that activity engagement is personally endorsed. The autonomous internalization comes from an intrinsic tendency of the self (Deci & Ryan, 1985, 2000, 2002; Ryan & Deci, 2003) and produces a motivational force to engage in the activity willingly (Vallerand, 1997; Vallerand, Fortier, & Guay, 1997). Individuals do not feel an uncontrollable urge to engage in the passionate activity, but rather freely choose to do so. Thus, when it comes to harmonious passion, behavioral engagement can be seen as flexible: People are able to decide when and when not to engage in the activity. For example, a teacher with a harmonious passion for teaching who is offered an opportunity to give a tutorial class at lunchtime might decide to decline the offer because such a task would reduce time that should be devoted to relaxing, reenergizing, and socializing with colleagues. The harmoniously passionate activity (e.g., teaching) can be seen as occupying a significant but not overpowering space in the person's identity and is in harmony with other aspects of the person's life (Vallerand et al., 2003). This should lead people with a harmonious passion to be able to fully concentrate on the task at hand and experience positive outcomes both during activity engagement (e.g., positive affect and flow) and after activity engagement (e.g., satisfaction or no guilt). Moreover, when prevented from taking part in their passionate activity, they should be able to focus their attention and energy on other tasks, without constantly ruminating about the passionate activity.

Conversely, obsessive passion results from a controlled internalization (Deci & Ryan, 2000) of the activity into one's identity. Such an internalization originates from intrapersonal or interpersonal pressure either because certain contingencies are attached to the activity, such as feelings of social acceptance or self-esteem, or because the sense of excitement derived from activity engagement becomes uncontrollable. For instance, although a teacher really enjoys his profession, he might experience an internal desire to teach because it is the only activity that might allow him to maintain a sense of self-worth. In such a case, teaching is no longer a truly volitional choice, but rather an activity this individual feels he *has* to do (e.g., to feel like a worthy person). People with an obsessive passion are controlled by their activity; it is as if they cannot help but to engage in it. Because the activity becomes out of one's control, it can take up disproportionate space in a person's life. This can lead one to neglect other life domains (e.g., family, friends, and leisure), thereby resulting in conflict in one's life. For

instance, a teacher with an obsessive passion for teaching might agree to partake in a committee organizing school activities—in addition to her regular class assignment—although such additional tasks would overload her and conflict with family activities. While attending the committee's meetings, she might be upset with herself for working instead of being at home with her husband and children. She might therefore have difficulties focusing on her task and may not experience positive affect while doing it. She might even feel guilty afterward. Thus, because of its controlled nature, it is proposed that obsessive passion can eventually lead to less adaptive outcomes. Although some benefits (e.g., improved teaching skills) might follow from this type of passionate involvement, so do harsh costs (e.g., personal and family problems and burnout).

Research has provided empirical support for several aspects of the passion conceptualization. First, the existence of two constructs, corresponding to harmonious and obsessive passion, has been supported by results from exploratory and confirmatory factor analyses with the Passion Scale (Rousseau, Vallerand, Ratelle, Mageau, & Provencher, 2002; Vallerand et al., 2003, Study 1; Vallerand et al., 2006, Study 1). Second, partial correlations (controlling for the correlation between the two types of passion) revealed that both harmonious and obsessive passion were positively associated with measures of activity valuation and measures of the activity being perceived as a passion, thereby providing support for the definition of passion. Empirical evidence has also shown that the two types of passion are associated with different affective experiences (Vallerand et al., 2003, Study 1). Thus, harmonious passion is positively associated with flow and positive emotions experienced during activity engagement, whereas obsessive passion is positively associated with negative emotions (e.g., shame) after engagement with the activity and when prevented from engaging in the activity altogether. Obsessive (but not harmonious) passion is also related to conflict with other aspects of one's life (Séguin-Lévesque, Laliberté, Pelletier, Blanchard, & Vallerand, 2003; Vallerand et al., 2003, Study 1), to rumination and negative affect when the person is prevented from engaging in the passionate activity (Ratelle, Vallerand, Mageau, Rousseau, & Provencher, 2004; Vallerand et al., 2003, Study 1), and to rigid persistence in ill-advised activities (Vallerand et al., 2003, Studies 3 and 4), eventually leading to chronic injuries in dancers (Rip, Fortin, & Vallerand, 2006) and to pathological gambling (Philippe & Vallerand, 2007; Ratelle et al., 2004).

It should be noted that support for these findings has been obtained in a variety of activities including work (Vallerand & Houliort, 2003), gambling (Mageau, Vallerand, Rousseau, Ratelle, & Provencher, 2005; Ratelle et al., 2004; Rousseau et al., 2002), Internet use (Séguin-Lévesque et al., 2003), sports (Vallerand et al., 2003, Study 2; Vallerand & Miquelon, 2007; Vallerand et al., 2006), and several types of recreational activities such as reading and playing music (Vallerand et al., 2003, Study 1).

### On Passion for Teaching: The Present Research

Passion appears to be an important concept to consider in education because the teaching profession requires teachers to invest time, energy, and their hearts in their teaching (Day, 2004). Passion is not only what attracts and keeps teachers in the profession (Nias, 1996), it is also what makes them constantly search for



more effective ways to reach their students and master their craft (Zehm & Kottler, 1993). Thus "passion is not an option. It is essential to high-quality teaching" (Day, 2004, p.3). Although several authors have emphasized the importance of passion in the teaching profession (Day, 2004; Elliott & Crosswell, 2001; Fried, 1995; Nias, 1996), the construct has not yet been the subject of extensive empirical research. The scientific study of passion in the teaching realm is especially relevant because paradoxical conceptions of passion have been proposed in the education literature. Thus, it appears that "passion can lead to enhanced vision [. . .] but it can also restrict wider vision and lead to the narrow pursuit of a passionately held conviction at the expense of other things" (Day, 2004, p. 11). The main purpose of the present research, therefore, was to use the recent theoretical framework of passion proposed by Vallerand et al. (2003) to provide a better understanding of the role of passion for teaching in various teacher and student outcomes.

School boards and administrators are increasingly concerned with teachers' professional satisfaction and burnout, two personal outcomes that play a crucial role with regards to teachers' intentions to stay committed (or not) to their jobs (Fore, Martin, & Bender, 2002; Shann, 1998). Although no empirical study has looked directly at the relationship between passion for teaching and professional outcomes, research on passion in other domains has clearly suggested that harmonious and obsessive passion should be differently associated with satisfaction and burnout in the teaching realm. Indeed, in previous studies, harmonious passion has been shown to lead to a number of adaptive outcomes, including enhanced subjective well-being and vitality (Rousseau & Vallerand, 2003, 2006; Vallerand et al., 2006, 2007, 2008), reduced depression and anxiety, positive affective experiences at work (Vallerand & Houliort, 2003), and positive emotions during and after engagement in the passionate activity (Vallerand et al., 2003, Study 1). Conversely, obsessive passion has been associated with maladaptive outcomes such as reduced work satisfaction (Vallerand & Houliort, 2003) and negative emotions both during and after engagement in the passionate activity (Vallerand et al., 2003, Study 1). Thus, one purpose of the present study was to replicate and extend past findings on the more adaptive outcomes derived from harmonious (vs. obsessive) passion with respect to work satisfaction and burnout.

Much has been written about the potential of passion for teaching to produce student benefits (e.g., Day, 2004; Fried, 1995). However, to the best of our knowledge, no empirical work has tested this relationship. Obsessive and harmonious passion should lead to positive student outcomes because enthusiasm, which is a key overt characteristic of passion (Selder & Paustian, 1989), has been shown to promote students' vitality (Patrick et al., 2000). Another purpose of this study was therefore to remedy the lack of research concerning the role that teachers' passion may play in students' adaptive classroom behaviors. An additional purpose was to use a cross-lag panel model to test for the directionality of effects between passion and outcomes. Specifically, such a design allowed us to determine whether changes in harmonious and obsessive passion predicted changes in outcomes (i.e., work satisfaction, burnout, and teacher-perceived student behaviors) over time or, conversely, whether changes in outcomes predicted changes in passion. A final and exploratory purpose of the present study was to assess the prevalence of passion in the teaching realm.

Overall, we made a number of predictions. First, we expected increases in harmonious passion to predict increases in work satisfaction over time and increases in obsessive passion either to predict losses of work satisfaction or to be unrelated to it. Second, we hypothesized that increases in harmonious passion would predict losses in burnout symptoms over time, whereas we expected increases in obsessive passion either to predict increases in burnout symptoms or to be unrelated to them. Third, we expected increases in both harmonious and obsessive passion to predict enhanced perceptions of positive student behaviors over time. Fourth, although we hypothesized that changes in harmonious and obsessive passion at Time 1 would predict changes in specific outcomes (as detailed above), we did not expect the reverse. That is, we did not expect changes in work satisfaction, burnout symptoms, and teacher-perceived student classroom behaviors to predict changes in both types of passions. Finally, we hypothesized that a majority of teachers would be passionate.

## Method

### Participants

Participants were 494 teachers (373 women, 119 men, and 2 who did not identify their gender) from French-Canadian schools in two school boards from the Quebec City area. Our sample consisted of 306 elementary teachers, 120 high school teachers, 20 teachers in adult education, 46 teachers in vocational and technical education, and 2 who did not specify their school level. Age ranged from 23 to 64 years ( $M = 43.07$  years,  $SD = 10.16$ ). The mean number of years of teaching experience was 15.82 years ( $SD = 10.33$ ). A total of 653 participants took part at Time 1; of these, 494 also took part at Time 2, yielding a response rate of 75.7%.<sup>1</sup>

### Procedure

This study was conducted over a 3 month-period and consisted of two data collections (i.e., March and June). At each phase of the study, teachers were asked to fill out a questionnaire and return it in a prestamped envelope. Follow-up telephone calls were made to increase the return rate.

### Instruments

**Passion.** The Passion Scale (Vallerand et al., 2003) contains two sections. The first assesses the extent to which people have a passion for an activity (i.e., teaching in the present study). The level of passion is measured with the mean of four criterion items that reflect the definition of passion. Specifically, participants are asked to report the extent to which they value the activity, devote time to it, love it, and view it as a passion (see appendix). These four items were highly intercorrelated in the present study: the

<sup>1</sup> Respondents who took part in both phases of the study were significantly older and had more years of experience as a teacher (as assessed at Time 1) than respondents who only took part in Time 1. However, controlling for those variables in the model did not change the pattern of results.



Table 1  
Means, Standard Deviations, and Correlations Involving the Latent Variables of the Model

Variable	M	SD	1	2	3	4	5	6	7	8	9	10
1. Harmonious passion (at T1)	4.92	1.02	—									
2. Obsessive passion (at T1)	2.65	1.04	-.14*	—								
3. Work satisfaction (at T1)	4.44	1.16	.73**	-.07	—							
4. Burnout (at T1)	1.97	1.00	-.63**	.38**	-.74**	—						
5. Positive student behaviors (at T1)	2.85	0.56	.32**	.02	.40**	-.36**	—					
6. Harmonious passion (at T2)	4.95	1.02	.80**	-.11*	.58**	-.50**	.26**	—				
7. Obsessive passion (at T2)	2.59	1.06	-.13**	.88**	-.06	.34**	.01	-.10*	—			
8. Work satisfaction (at T2)	4.50	1.20	.71**	-.08	.83**	-.63**	.34**	.73**	-.09*	—		
9. Burnout (at T2)	2.00	1.04	-.62**	.31**	-.66**	.85**	-.32**	-.60**	.32**	-.70**	—	
10. Positive student behaviors (at T2)	2.90	0.59	.36**	.10*	.38**	-.31**	.78**	.31**	.09*	.38**	-.36**	—

Note.  $N = 494$ ; all scales were assessed on a 7-point scale except for teacher-perceived student behaviors, which were assessed on a 4-point scale. T1 = Time 1; T2 = Time 2.

\*  $p < .05$ . \*\*  $p < .01$ .

Cronbach's alphas were .79 and .78 at Times 1 and 2, respectively. The second section of the Passion Scale assesses harmonious and obsessive passions with two six-item subscales. In the present study, harmonious and obsessive passions were found to be unrelated ( $r_s = .01$  and  $.00$ , at Times 1 and 2, respectively).<sup>2</sup> A sample item for harmonious passion is "My job as a teacher is in harmony with the other activities in my life"; a sample item for obsessive passion is "I have almost an obsessive feeling for my job as a teacher." The Cronbach's alpha values for these two subscales were .87 and .76, respectively, at Time 1 and .87 and .80, respectively, at Time 2. Responses to all items were scored on a 7-point Likert scale ranging from 1 (*do not agree at all*) to 7 (*very strongly agree*). Much support exists for the validity and the reliability of the Passion Scale (see Rousseau et al., 2002; Vallerand et al., 2003, 2006). In this study, a confirmatory factor analysis of the Passion Scale yielded an acceptable fit to the data,  $\chi^2(23, N = 494) = 138.12, p < .001$ , comparative fit index = .96, normed fit index = .95, standardized root-mean-square residual = .08, root-mean-square error of approximation = .10, confidence interval = [.09–.12].

**Work satisfaction.** We assessed work satisfaction with the French-Canadian version (Blais, Vallerand, Pelletier, & Brière, 1989) of the Satisfaction With Life Scale (Diener, Emmons, Larsen, & Griffin, 1985), adapted to the worklife for the purpose of this study. The instrument consists of five items that are measured on a 7-point Likert scale, ranging from 1 (*strongly disagree*) to 7 (*strongly agree*). A sample item is "In most ways my job is close to my ideal." The Cronbach's alphas in this study were .88 and .90 at Times 1 and 2, respectively.

**Burnout.** Burnout was assessed using the French-Canadian version (Dion & Tessier, 1994) of the Maslach Burnout Inventory (Maslach & Jackson, 1986). The psychometric properties of the French-Canadian version of the Maslach Burnout Inventory are similar to those of the original version (see Dion & Tessier, 1994). The instrument measures three dimensions of burnout: emotional exhaustion, depersonalization, and personal accomplishment. However, in line with past research (Coders & Dougherty, 1993; Shirom, 2003), we used only the first two dimensions—considered the "core of burnout"—in the present study. Emotional exhaustion (nine items) is characterized by a lack of energy and a feeling that

one's emotional resources are used up (e.g., "I feel emotionally drained from my work"). Depersonalization (five items) refers to a negative, callous, and detached attitude toward the people with whom one works (e.g., "I feel I treat some students as if they were impersonal objects"). Responses were scored on a 7-point Likert scale varying from 1 (*never*) to 7 (*everyday*). The Cronbach's alpha values were .70 and .77 for depersonalization and .92 and .93 for emotional exhaustion, at Times 1 and 2, respectively.

**Teacher-perceived student behaviors.** We used three items from the French-Canadian version (Fernet & Senécal, 2004) of the Pupil Behavior Patterns Scale (Friedman, 1995) to assess teachers' perceptions of student behavior patterns. Responses were scored on a 4-point Likert scale varying from 1 (*never*) to 4 (*very often*). A sample item was "Students in my class are cooperative and enthusiastic." High scores are indicative of adaptive student classroom behaviors. The Cronbach's alphas in this study were .73 and .75 at Times 1 and 2, respectively.

## Results

Means, standard deviations, and correlations of the model variables are presented in Table 1. We performed all structural equation modeling analyses on a raw data file using maximum likelihood estimation procedure (EQS version 6.1; Bentler, 1993). The model tested in the present study was composed of 10 latent variables: 5 exogenous variables (i.e., harmonious passion, obsessive passion, work satisfaction, burnout, and student behaviors at Time 1) and 5 endogenous variables (i.e., harmonious passion, obsessive passion, work satisfaction, burnout, and student behaviors at Time 2). As shown in Figure 1, each latent variable had between two and five indicators. The three items of the Pupil Behavior Patterns Scale were used as the indicators of the student behaviors factor, and the five items of the Work Satisfaction Scale were used as the indicators of the work satisfaction latent variable.

<sup>2</sup> Correlations between harmonious and obsessive passions have been found to vary across studies. This would suggest that whether the two types of passion are orthogonal or not may be a function of the type of activity at hand.

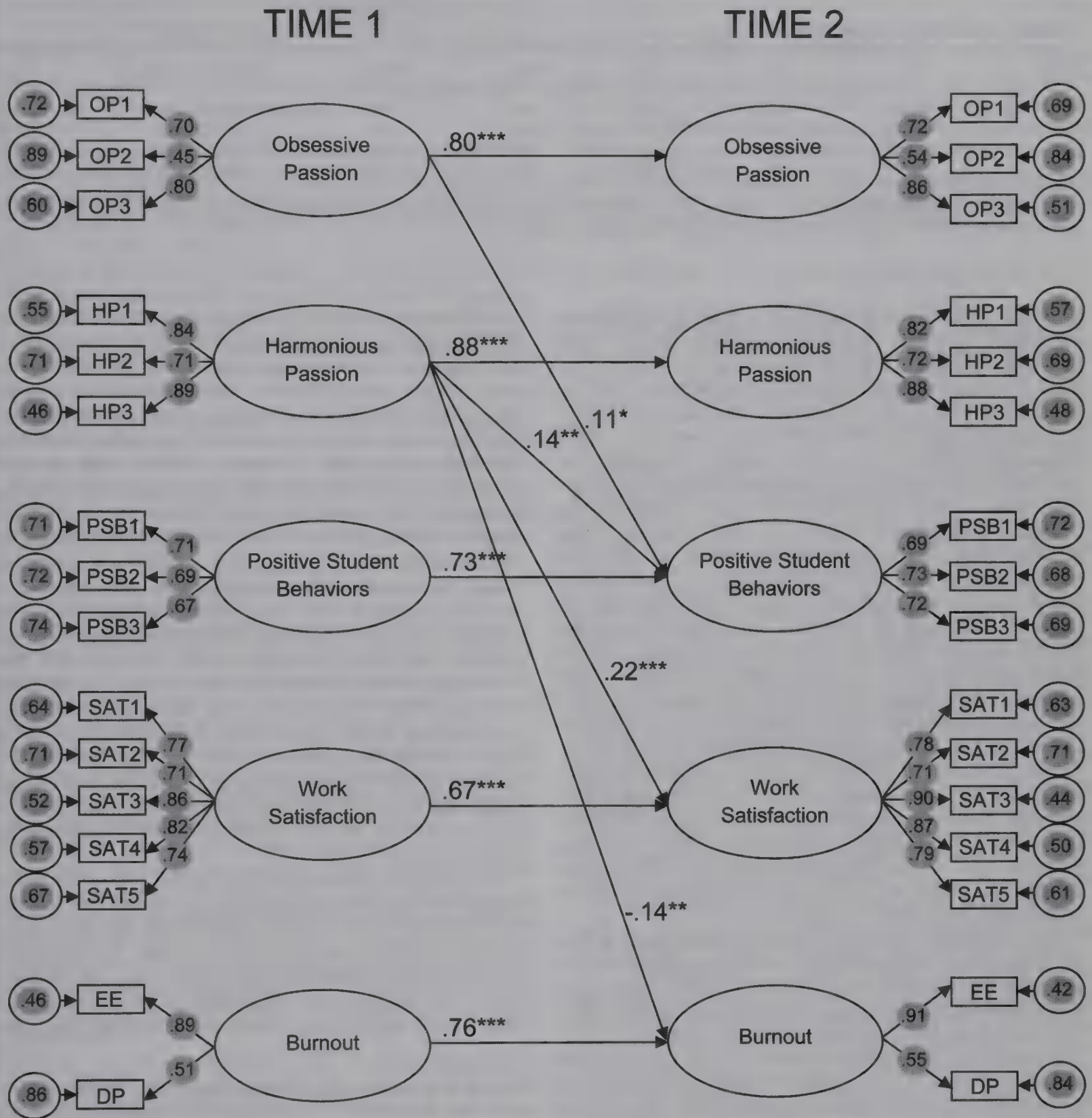


Figure 1. Results of the structural equation modeling analyses.  $N = 494$ . \* $p < .05$ . \*\* $p < .01$ . \*\*\* $p < .001$ .

Because the variables contained too many items, parcels were used as indicators of the two types of passion as well as the burnout latent variable. For harmonious and obsessive passion, we computed parcels by aggregating Items 1 and 2 from their respective subscale into Parcel 1, Items 3 and 4 into Parcel 2, and Items 5 and 6 into Parcel 3. Parcels were also used for the burnout variable, in which we aggregated items from the Emotional Exhaustion subscale into Parcel 1 and items from the Depersonalization subscale

into Parcel 2. Following suggestions by Marsh and Hau (1996), we allowed each of the 16 indicators at Time 1 to covary with its equivalent at Time 2 to prevent inflated estimates of stability. We estimated the covariances between the five latent constructs at Time 1 and the covariances between the disturbance terms at Time 2. For clarity concerns, we omitted covariances in Figure 1.

To test the hypotheses of the present study, we tested a first model. A total of 11 paths were specified: one between each



Table 2  
*Goodness-of-Fit Indices of the Four Models*

Model	$\chi^2$	df	CFI	NFI	SRMR	RMSEA and CI	AIC
1	949.67	417	.95	.91	.08	.05 (.05-.06)	115.67
2	952.43	419	.95	.91	.08	.05 (.05-.06)	114.43
3	972.45	417	.95	.91	.08	.05 (.05-.06)	138.45
4	941.79	411	.95	.91	.08	.05 (.05-.06)	119.79

*Note.* CFI = comparative fit index; NFI = normed fit index; SRMR = standardized root-mean-square residual; RMSEA = root-mean-square error of approximation; CI = confidence interval; AIC = Akaike's information criterion.

variable at Time 1 and its equivalent at Time 2, one between harmonious passion at Time 1 and each outcome (i.e., work satisfaction, burnout, and teacher-perceived student behaviors) at Time 2, and one between obsessive passion at Time 1 and the outcomes at Time 2. The results showed that the model had an acceptable fit to the data,  $\chi^2(417, N = 494) = 949.67, p < .001$ , comparative fit index = .95, normed fit index = .91, standardized root-mean-square residual = .08, root-mean-square error of approximation = .05, confidence interval = [.05-.06]. Because the chi-square is sensitive to sample size, some researchers have suggested using the normed chi-square, which is the chi-square value divided by the degrees of freedom (Kline, 2005). Bollen (1989) suggested that a normed chi-square value of less than 3.0 indicates a reasonable fit to the data. In the present study, the normed chi-square value was 2.28 (949.67/417). Results revealed that all paths but two (i.e., those from obsessive passion to work satisfaction and burnout) were significant. We conducted the analyses again, omitting the two nonsignificant paths. The model fit indices remained virtually unchanged, as shown in Table 2 (see Model 2). To compare the two models, we used Akaike's information criterion (AIC) statistic. Within a set of competing nonhierarchical models, the one with the lowest AIC value is preferred (Kline, 2005). The original model yielded an AIC of 115.67. The model in which two paths were omitted (i.e., Model 2) was kept because it yielded an AIC of 114.43, which indicates that it was slightly more parsimonious than Model 1.

The standardized solutions for Model 2 are presented in Figure 1. Each latent variable at Time 1 was strongly and positively associated with its equivalent at Time 2 ( $\beta$ s ranging from 0.67 to 0.88), suggesting that the constructs are relatively stable over time. Although a large part of the variance of each outcome at Time 2 was explained by the equivalent outcome at Time 1, a significant part of the variance of those outcomes was nevertheless explained by either one (for work satisfaction and burnout) or both (for student behaviors) types of passion at Time 1. More specifically, increases in harmonious passion predicted increases in work satisfaction ( $\beta = 0.22$ ) and increases in teacher-perceived adaptive student behaviors ( $\beta = 0.14$ ), as well as decreases in burnout ( $\beta = -0.14$ ) that took place over the 3-month period. We also found that increases in obsessive passion predicted an increased perception of adaptive student behaviors ( $\beta = 0.11$ ) from Time 1 to Time 2.<sup>3</sup>

We tested two alternative models. The first alternative model (i.e., Model 3) considered in this study was one in which the paths from outcomes (i.e., work satisfaction, burnout, and student behaviors) at Time 1 to harmonious and obsessive passions at Time

2 were estimated in addition to the paths from all five variables at Time 1 to their equivalents at Time 2. We tested such a model to help determine the directionality of effects (i.e., whether harmonious and obsessive passions are the antecedents and not the consequences of work satisfaction, burnout, and perceived student behaviors). As can be seen in Table 2, this model resulted in increased chi-square and AIC values. Moreover, results revealed the absence of significant paths from the outcomes at Time 1 to the two types of passion at Time 2. Model 3 was therefore rejected. Finally, we tested another alternative model involving reciprocal effects (i.e., Model 4). In such a model, the paths from passions at Time 1 to outcomes at Time 2 and the paths from outcomes at Time 1 to passions at Time 2 were estimated simultaneously. The model fit indices are shown in Table 2. Model 4 provided no additional information in comparison to the previous models; that is, the paths from harmonious passion at Time 1 to the three outcomes at Time 2 were significant, as was the path from obsessive passion at Time 1 to teacher-perceived student behaviors at Time 2, whereas no significant relations were found between the outcomes at Time 1 and the passions at Time 2. For those reasons and because the model AIC value was higher than that of Model 2 (119.79 > 114.43), the latter was preferred, thereby supporting the proposed model.

As a final and exploratory purpose, we had proposed to assess the prevalence of passion for teaching in our sample. We used the four criteria of passion (i.e., activity valuation, time investment, love for the activity, and activity being a passion; Vallerand et al., 2003) to differentiate between the nonpassionate and the passionate teachers. Teachers with a mean score on the sum of the four criteria at the midpoint (4) or above on the response scale were

<sup>3</sup> The model was also tested with estimated missing values at Time 2 ( $N = 653$ ) using the full-information maximum likelihood method. The same results were obtained. Furthermore, we tested a Time-1-only model and a Time-2-only model. In the Time-1-only model, harmonious passion was positively associated with satisfaction ( $\beta = 0.78$ ) and student behavior ( $\beta = 0.38$ ) and negatively associated with burnout ( $\beta = -0.63$ ), whereas obsessive passion was positively associated with burnout ( $\beta = 0.28$ ). In the Time-2-only model, harmonious passion was positively associated with satisfaction ( $\beta = 0.81$ ) and student behavior ( $\beta = 0.38$ ) and negatively associated with burnout ( $\beta = -0.65$ ), and obsessive passion was positively associated with burnout ( $\beta = 0.27$ ) and student behavior ( $\beta = 0.14$ ). Both models yielded acceptable fits to the data. Thus, both Time-1-only and Time-2-only models yielded the same basic results as the model involving both time points. We feel that these results provide additional support for our model.

classified as passionate. Results showed that 93.1% of the teachers in the present study were at least moderately passionate about their teaching.<sup>4</sup>

## Discussion

The general purpose of the present research was to proceed to the empirical study of passion in the teaching realm to achieve a better understanding of the intrapersonal and interpersonal outcomes associated with it. Specifically, we sought to shed light on the role of changes in harmonious and obsessive passions in the changes in work satisfaction, burnout, and teacher-perceived positive student behaviors that take place over time. We expected increases in harmonious passion to predict an augmentation of work satisfaction and a reduction of burnout symptoms over time. Conversely, we expected increases in obsessive passion either to predict a reduction of work satisfaction and an augmentation of burnout symptoms over time or to be unrelated to those outcomes. Moreover, we expected increases in both harmonious and obsessive passions to predict increases in teacher-perceived positive student behaviors over time. Finally, we hypothesized that most teachers would be passionate toward their job. Overall, the results of the present research provided support for all these hypotheses.

### *On Passion for Teaching and Intrapersonal Outcomes*

An important conclusion that may be drawn from the present findings has to do with the relationships between passion and intrapersonal outcomes. In line with our hypotheses, increases in harmonious passion were shown to predict increases in job satisfaction and decreases in burnout symptoms over time, and increases in obsessive passion were shown to be unrelated to such outcomes. These findings are particularly interesting because they shed some light on the following inconsistency: Passion seems to lead some teachers to fully enjoy themselves in their jobs, whereas others do not appear to derive such positive benefits (Day, 2004; Nias, 1989). Therefore, our results support the assumption that the quality of outcomes derived from one's teaching depends on the type of passion held by the teacher. It would thus appear that involvement in an activity central to one's life (such as teaching) contributes to one's personal well-being to the extent that harmonious passion underlies such involvement. Conversely, high activity involvement triggered by obsessive passion may not afford similar benefits.

Future research is needed to identify the psychological processes through which harmonious passion is linked to reduced burnout and enhanced job satisfaction. Past research on passion has shown that during task engagement, harmonious passion is associated with positive affect, whereas obsessive passion is unrelated to positive affect and can even predict negative affect (Mageau et al., 2005; Vallerand et al., 2003, Study 1). Thus, one plausible explanation for the effect of harmonious passion on the intrapersonal outcomes may have to do with the fact that engaging in an activity (such as teaching) out of harmonious passion leads to the cumulative experience of positive affect, which over time translates into increased satisfaction and subjective well-being. This would be consistent with research by Fredrickson and Joiner (2002) that has shown the existence of an upward spiral, whereby positive affect leads to higher levels of subjective well-being that

lead to subsequent experiences of positive affect, and so on. Such a spiral may be triggered by the impact of harmonious passion on positive affect. Similarly, a downward spiral involving negative affect and subjective well-being might also exist and may be induced by obsessive passion. In fact, research with elderly individuals by Rousseau and Vallerand (2003, 2008) has suggested that harmonious passion could lead to positive affect that, in turn, leads to increased subjective well-being. The role of affect as a mediator between passion for teaching and intrapersonal outcomes should be more fully examined in future research.

### *On Passion for Teaching and Interpersonal Outcomes*

The present research provides valuable information on the role of passion for teaching in teacher-perceived student outcomes. Thus, increases in both types of passion were shown to lead to increases in teacher-perceived adaptive student classroom behaviors over a 3-month period. Although this result is in line with many authors' claims that teachers who display a passion for their teaching have a positive impact on their students (Day, 2004; Fried, 1995), this study was the first to test this sequence within a theoretical framework. The finding that harmonious and obsessive passion have a similar and positive impact on others is interesting because the two types of passion are fueled by quite different processes. Thus, on one hand, obsessive passion originates largely from ego-invested structures within the person (Hodgins & Knee, 2002), which lead the individual to be defensive and to not open up to the world. On the other hand, harmonious passion emanates from the authentic integrating self (Deci & Ryan, 2000; Hodgins & Knee, 2002) and allows the individual to fully open up to various experiences. We suggest that students might perceive the vitality, intensity, and enthusiasm (Patrick et al., 2000) common to harmonious and obsessive passion because those are visible signs of passion. However, we posit that students are unaware of the processes underlying the two types of passion and, therefore, would not be expected to distinguish an obsessively passionate from a harmoniously passionate teacher. This would explain why increases in teacher-perceived positive student behaviors were positively predicted by increases in both harmonious and obsessive passions, although teachers themselves did not derive the same personal benefits from the two types of passion. Further examination of the interpersonal impact of passion is clearly needed. In fact, although the type of passion a teacher experiences does not seem to make a difference for students, it may not be the case for other individuals surrounding passionate teachers. Indeed, research by Séguin-Lévesque et al. (2003) revealed that obsessive passion toward the Internet was positively associated with conflict in the couple relationship but negatively related to dyadic adjustment. In contrast, harmonious passion toward the Internet was associated with less conflict and greater dyadic adjustment. Future research should investigate whether harmoniously and obsessively passionate teachers might have different impacts on significant others such as spouses and colleagues.

Another issue related to the interpersonal domain would be whether the two types of passion lead to different interpersonal

<sup>4</sup> Removing the nonpassionate teachers ( $N = 34$ ) from the analyses did not change the pattern of results. Therefore, these participants were not excluded.



effects as a function of the kind of activities at hand. Although there might be no visible signs allowing others to distinguish teachers' harmonious passion from their obsessive passion in the classroom, it could be otherwise in other settings. For example, in the physical activity realm, a teacher with an obsessive passion might be seen as more aggressive and competitive by people surrounding him or her than one with a harmonious passion. These different types of passion might lead to different effects on students. Such a hypothesis constitutes an interesting research avenue to explore.

### *Implications for the Dualistic Model of Passion*

Another implication from the present results pertains to the support they provide for the dualistic model of passion. Our findings first document the differential influences of harmonious and obsessive passion on the intrapersonal level. They showed that although obsessive passion is unrelated to well-being benefits, harmonious passion can make a significant difference in teachers' lives by boosting job satisfaction and decreasing the risks of burnout over time. These results are congruent with those from previous studies that demonstrated that harmonious passion is associated with more adaptive outcomes than is obsessive passion (Mageau et al., 2005; Ratelle et al., 2004; Rip et al., 2006; Rousseau & Vallerand, 2003, 2008; Vallerand et al., 2003, 2006). Although results showed that changes in obsessive passion did not predict changes in burnout symptoms, it is worth noting that there was an association between these two variables when the measurement times were considered separately ( $\beta = 0.28$  at Time 1 and  $\beta = 0.27$  at Time 2).

The present study also yielded important findings with respect to interpersonal influences of passion (Vallerand et al., 2003). Results revealed that although divergent at the intrapersonal level, outcomes associated with harmonious and obsessive passions for teaching converge at the interpersonal level. Specifically, results showed that both types of passion foster adaptive student behaviors in the classroom, as perceived by the teacher. They also suggest that the mere fact of being passionate for one's teaching would be enough to instill adaptive student classroom behavior. Future research is needed to identify the nature of the psychological processes that mediate the positive effects of passion on such student adaptive behavior. For instance, does teacher's enthusiasm (Patrick et al., 2000) induced by passion represent a crucial mediator of such effects? This hypothesis should be tested in future research.

Results also provide valuable information about the directionality of effects between passion and outcomes. Thus, harmonious and obsessive passions at Time 1 were significantly associated with the outcomes at Time 2, whereas the reverse (i.e., the association between the outcomes at Time 1 and the passions at Time 2) was not found. These findings support the fact that passion might be more of an antecedent than a consequence of work satisfaction, burnout, and perceptions of adaptive student behaviors. These findings are in line with past research that has shown that passion is implicated in a variety of personal and professional outcomes such as subjective well-being and vitality (Rousseau & Vallerand, 2003, 2008; Vallerand et al., 2006) and reduced depression, reduced anxiety, and positive affective experiences at work (Vallerand & Houlfort, 2003). Thus, additional research over a

longer period of time (e.g., a whole academic year) is needed to replicate the present findings.

Finally, much work remains to be done with respect to understanding the processes underlying the development of passion. Why is it that some teachers develop an obsessive passion toward teaching and others take the more beneficial road of harmonious passion? Work by Mageau et al. (in press) has suggested that an autonomy-supportive social environment (i.e., where choices are offered, initiatives are supported, and control is minimized) promotes the emergence of harmonious passion. Further research to clarify the antecedents of passion for teaching appears in order. Such research could result in important theoretical and applied advances, eventually leading to the creation of programs designed to facilitate the development of teachers' harmonious passion, which would be beneficial for teachers themselves and for their students.

### *On the Prevalence of Passion in the Teaching Realm*

An implication of the present research worth noting is that it empirically showed that passion for teaching appears to be highly prevalent. The hypothesis that a majority of teachers would be passionate toward their job was largely confirmed. Indeed, 93.1% of the teachers in the present sample met the criteria for at least a moderate level of passion toward teaching (Vallerand et al., 2003). These results confirm the widespread notion that teachers are passionate about their job (Day, 2004; Elliott & Crosswell, 2001; Fried, 1995; Nias, 1996) and highlight the relevance of studying passion in the teaching realm. Thus, our study fills an empirical gap by being the first to support the widely accepted claim that teachers are passionate. However, results about the prevalence of passion for teaching are limited to our sample. Thus, these results cannot be generalized to the broader population of teachers. Further research with a sample representative of the teaching population is clearly needed on this issue.

### *Limitations*

Although the results from the present study are consistent with a causal interpretation, the data are correlational in nature, and therefore, definitive conclusions about causality are not warranted. Future research using experimental designs should be used to replicate and confirm the proposed model. Another limitation to consider is that the coefficients of change—although statistically significant—were low ( $\beta$ s ranging from 0.11 to 0.22). This might be because of the short interval (i.e., 3 months) between the two measurement times that did not allow for much change to take place. This interpretation is supported by the stability coefficients that were rather high (see Figure 1). Therefore, interpretation of the present results should be done with caution. Clearly, subsequent research should use a longitudinal design extending over an entire school year to see whether the present results are replicated and whether higher coefficients of change might be obtained. Finally, we should underscore that the measure assessing student positive behaviors was completed by teachers. Thus, although the results showed that the more passionate the teachers, the more they perceived students to display adaptive behaviors, we cannot conclude that such a finding reflects actual positive student behaviors. Indeed, it is possible that changes in student behavior were because

of changes in teachers' perceptions. Therefore, in future research, it would be important for student behaviors to be reported by the students themselves or assessed by a third party (e.g., through videos), for objectivity concerns.

### Conclusion

In sum, the present findings provide interesting answers about the presence and role of passion in the teaching realm. Because harmonious passion promotes positive intrapersonal outcomes and relates to perceived student outcomes, future research is clearly needed to more completely understand the determinants of harmonious passion for teaching. Furthermore, interventions designed to promote a harmonious passion should also be the focus of future research. This would help teachers personally derive the best out of their teaching involvement while sustaining students' interest toward education.

### References

- Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, 63, 596–612.
- Bentler, P. M. (1993). *EQS: Structural equation program manual*. Los Angeles: BMDP Statistical Software.
- Blais, M. R., Vallerand, R. J., Pelletier, L. G., & Brière, N. M. (1989). L'échelle de satisfaction de vie: Validation canadienne-française du "Satisfaction With Life Scale" [French-Canadian validation of the Satisfaction With Life Scale]. *Canadian Journal of Behavioural Science*, 21, 210–223.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Codgers, C. L., & Dougherty, T. W. (1993). A review and integration of research on job burnout. *Academy of Management Review*, 18, 621–656.
- Coulehan, J. L. (2002). Being a physician: The physician's power. In M. B. Mengel, W. L. Holleman, & S. A. Fields (Eds.), *Fundamentals of clinical practice* (2nd ed., pp. 73–97). New York: Kluwer Academic/Plenum.
- Csikszentmihalyi, M., Rathunde, K., & Whalen, S. (1993). *Talented teenagers: The roots of success & failure*. New York: Cambridge University Press.
- Day, C. (2004). *A passion for teaching*. London: Routledge Falmer.
- Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. New York: Plenum.
- Deci, E. L., & Ryan, R. M. (2000). The "what" and the "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, 11, 227–268.
- Deci, E. L., & Ryan, R. M. (2002). *Handbook of self-determination research*. Rochester, NY: University of Rochester Press.
- Diener, E., Emmons, R. A., Larsen, R. J., & Griffin, S. (1985). The Satisfaction With Life Scale. *Journal of Personality Assessment*, 49, 71–76.
- Dion, G., & Tessier, R. (1994). Validation de la traduction de l'Inventaire d'épuisement professionnel de Maslach et Jackson [Validation and translation of the burnout inventory of Maslach and Jackson]. *Canadian Journal of Behavioural Science*, 26, 210–227.
- Dlugos, R. F., & Friedlander, M. L. (2001). Passionately committed psychotherapists: A qualitative study of their experience. *Professional Psychology: Research and Practice*, 32, 298–304.
- Elliott, B., & Crosswell, L. (2001). *Commitment to teaching: Australian perspectives on the interplays of the professional and the personal in teachers' lives*. Paper presented at the International Symposium on Teacher Commitment at the European Conference on Educational Research, Lille, France.
- Fernet, C., & Senécal, C. (2004). *Adaptation Canadienne-Française du Pupil Behavior Patterns scale*. Unpublished manuscript, Université Laval, Québec, Canada.
- Fore, C., Martin, C., & Bender, W. (2002). Teacher burnout in special education: The causes and the recommended solutions. *High School Journal*, 86, 36–45.
- Fredrickson, B. L., & Joiner, T. (2002). Positive emotions trigger upward spirals toward emotional well-being. *Psychological Science*, 13, 172–175.
- Freudenberger, H. J. (1985). Impaired clinicians: Coping with burnout. In P. A. Keller (Ed.), *Innovations in clinical practice: A source book 3*. Sarasota, FL: Professional Resource Exchange.
- Fried, R. L. (1995). *The passionate teacher: A practical guide*. Boston: Beacon Press.
- Friedman, I. A. (1995). Student behavior patterns contributing to teacher burnout. *Journal of Educational Research*, 88, 281–289.
- Grosch, W. N., & Olsen, D. C. (1994). *When helping starts to hurt: A new look at burnout among psychotherapists*. New York: W. W. Norton.
- Hodgins, H. S., & Knee, R. (2002). The integrating self and conscious experience. In E. L. Deci & R. M. Ryan (Eds.), *Handbook of self-determination research* (pp. 87–100). Rochester, NY: University of Rochester Press.
- Ingersoll, R. M. (2001). Teacher turnover and teacher shortages: An organizational analysis. *American Educational Research Journal*, 38, 499–534.
- Kline, R. B. (2005). *Principles and practice of structural equation modeling*. New York: Guilford Press.
- Mageau, G. A., Vallerand, R. J., Charest, J., Salvy, S.-J., Lacaille, N., Bouffard, T., & Koestner, R. (in press). On the development of harmonious and obsessive passion: The role of autonomy support, activity specialization and identification with the activity. *Journal of Personality*.
- Mageau, G. A., Vallerand, R. J., Rousseau, F. L., Ratelle, C. F., & Provencher, P. J. (2005). Passion and gambling: Investigating the divergent affective and cognitive consequences of gambling. *Journal of Applied Social Psychology*, 35, 100–118.
- Marcil, M. (1991). *Stress et burnout*. Montréal: ACSM.
- Marsh, H. W., & Hau, K. T. (1996). Assessing goodness of fit: When parsimony is undesirable. *Journal of Experimental Education*, 64, 364–390.
- Maslach, C., & Jackson, S. E. (1986). *Maslach Burnout Inventory: Manual* (2nd ed.). Palo Alto, CA: Consulting Psychologists Press.
- Nias, J. (1989). *Primary teachers talking: A study of teaching and work*. London: Routledge.
- Nias, J. (1996). Thinking about feeling: The emotions in teaching. *Cambridge Journal of Education*, 26, 293–306.
- Patrick, B. C., Hisley, J., Kempler, T., & College, G. (2000). "What's everybody so excited about?: The effects of teacher enthusiasm on student intrinsic motivation and vitality. *Journal of Experimental Education*, 68, 217–236.
- Philippe, F., & Vallerand, R. J. (2007). Prevalence rates of gambling problems in Montreal, Canada: A look at old adults and the role of passion. *Journal of Gambling Studies*, 23, 275–283.
- Ratelle, C. F., Vallerand, R. J., Mageau, G. A., Rousseau, F. L., & Provencher, P. J. (2004). When passion leads to pathology: A look at gambling. *Journal of Gambling Studies*, 20, 105–119.
- Rip, B., Fortin, S., & Vallerand, R. J. (2006). The relationship between passion and injury in dance students. *Journal of Dance Medicine & Science*, 10, 14–20.
- Rousseau, F. L., & Vallerand, R. J. (2003). Le rôle de la passion dans le bien-être subjectif des aînés [The role of passion in the subjective



- well-being of the elderly]. *Revue Québécoise de Psychologie*, 24, 197–211.
- Rousseau, F. L., & Vallerand, R. J. (2008). An examination of the relationship between passion and subjective well-being in older adults. *International Journal of Aging and Human Development*, 66, 195–211.
- Rousseau, F. L., Vallerand, R. J., Ratelle, C. F., Mageau, G. A., & Provencher, P. J. (2002). Passion and gambling: On the validation of the Gambling Passion Scale (GPS). *Journal of Gambling Studies*, 18, 45–66.
- Ryan, R. M., & Deci, E. L. (2003). On assimilating identities to the self: A self-determination theory perspective on internalization and integrity within cultures. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (pp. 255–273). New York: Guilford Press.
- Séguin-Lévesque, C., Laliberté, M.-L. N., Pelletier, L. G., Blanchard, C., & Vallerand, R. J. (2003). Harmonious and obsessive passion for the Internet: Their associations with the couples' relationships. *Journal of Applied Social Psychology*, 33, 197–221.
- Selder, F. E., & Paustian, A. (1989). Burnout: Absence of vision. *Loss, Grief and Care*, 3, 73–93.
- Shann, M. H. (1998). Professional commitment and satisfaction among teachers in urban middle schools. *Journal of Educational Research*, 9, 67–73.
- Shirom, A. (2003). Job-related burnout: A review. In J. C. Quick & L. Tetrick (Eds.), *Handbook of occupational health psychology* (pp. 245–264). Washington, DC: American Psychological Association.
- Vallerand, R. J. (1997). Toward a hierarchical model of intrinsic and extrinsic motivation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (pp. 271–360). New York: Academic Press.
- Vallerand, R. J., Blanchard, C., Mageau, G. A., Koestner, R., Ratelle, C. F., Leonard, M., et al. (2003). *Les passions de l'âme: On obsessive and harmonious passion*. *Journal of Personality and Social Psychology*, 85, 756–767.
- Vallerand, R. J., Fortier, M. S., & Guay, F. (1997). Self-determination and persistence in a real-life setting: Toward a motivational model of high school dropout. *Journal of Personality and Social Psychology*, 72, 1161–1176.
- Vallerand, R. J., & Houliort, N. (2003). Passion at work: Toward a new conceptualization. In D. Skarlicki, S. Gilliland, & D. Steiner (Eds.), *Social issues in management* (pp. 175–204). Greenwich, CT: Information Age.
- Vallerand, R. J., Mageau, G. A., Elliot, A., Dumais, A., Demers, M.-A., & Rousseau, F. L. (2008). Passion and performance attainment in sport. *Psychology of Sport & Exercise*, 9, 373–392.
- Vallerand, R. J., & Miquelon, P. (2007). Passion in sport: Theory, research, and applications. In D. Lavallée & S. Jowett (Eds.), *Social psychology in sport* (pp. 249–263). Champaign, IL: Human Kinetics.
- Vallerand, R. J., Rousseau, F. L., Grouzet, F. M. E., Dumais, A., Grenier, S., & Blanchard, C. M. (2006). Passion in sport: A look at determinants and affective experiences. *Journal of Sport and Exercise Psychology*, 28, 454–478.
- Vallerand, R. J., Salvy, S.-J., Mageau, G. A., Denis, P., Grouzet, F. M. E., & Blanchard, C. B. (2007). On the role of passion in performance. *Journal of Personality*, 75, 505–533.
- Wilhelm, K., Dewhurst-Savellis, J., & Parker, G. (2000). Teacher stress? An analysis of why teachers leave and why they stay. *Teachers and Teaching: Theory and Practice*, 6, 291–304.
- Zehm, S. J., & Kottler, J. A. (1993). *On being a teacher: The human dimension*. Thousand Oaks, CA: Corwin Press.

## Appendix

### The Passion Scale—Adapted for Teaching

1. I spend a lot of time doing my job as a teacher.
2. I like my job as a teacher.
3. My job as a teacher is important for me.
4. My job as a teacher is a passion for me.
5. My job as a teacher is in harmony with the other activities in my life.
6. I have difficulties controlling my urge to do my job as a teacher.
7. The new things that I discover doing my job as a teacher allow me to appreciate it even more.
8. I have almost an obsessive feeling for my job as a teacher.
9. My job as a teacher reflects the qualities I like about myself.
10. My job as a teacher allows me to live a variety of experiences.
11. My job as a teacher is the only thing that really turns me on.
12. My job as a teacher is well integrated in my life.
13. If I could, I would only do my job as a teacher.

14. My job as a teacher is in harmony with other things that are part of me.
15. My job as a teacher is so exciting that I sometimes lose control over it.
16. I have the impression that my job as a teacher controls me.

## Key for the Passion Scale

- # 1-4, Passion Criteria
- # 5, 7, 9, 10, 12, 14, Harmonious Passion
- # 6, 8, 11, 13, 15, 16, Obsessive Passion

Received July 10, 2007

Revision received February 12, 2008

Accepted April 23, 2008 ■



## AMERICAN PSYCHOLOGICAL ASSOCIATION

### SUBSCRIPTION CLAIMS INFORMATION

Today's Date: \_\_\_\_\_

We provide this form to assist members, institutions, and nonmember individuals with any subscription problems. With the appropriate information we can begin a resolution. If you use the services of an agent, please do **NOT** duplicate claims through them and directly to us. **PLEASE PRINT CLEARLY AND IN INK IF POSSIBLE.**

PRINT FULL NAME OR KEY NAME OF INSTITUTION \_\_\_\_\_

MEMBER OR CUSTOMER NUMBER (MAY BE FOUND ON ANY PAST ISSUE LABEL) \_\_\_\_\_

ADDRESS \_\_\_\_\_

DATE YOUR ORDER WAS MAILED (OR PHONED) \_\_\_\_\_

CITY \_\_\_\_\_ STATE/COUNTRY \_\_\_\_\_ ZIP \_\_\_\_\_

\_\_\_\_ PREPAID \_\_\_\_ CHECK \_\_\_\_ CHARGE

CHECK/CARD CLEARED DATE: \_\_\_\_\_

YOUR NAME AND PHONE NUMBER \_\_\_\_\_

(If possible, send a copy, front and back, of your cancelled check to help us in our research of your claim.)

ISSUES: \_\_\_\_ MISSING \_\_\_\_ DAMAGED

TITLE \_\_\_\_\_

VOLUME OR YEAR \_\_\_\_\_

NUMBER OR MONTH \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

*Thank you. Once a claim is received and resolved, delivery of replacement issues routinely takes 4-6 weeks.*

(TO BE FILLED OUT BY APA STAFF)

DATE RECEIVED: \_\_\_\_\_

DATE OF ACTION: \_\_\_\_\_

ACTION TAKEN: \_\_\_\_\_

INV. NO. &amp; DATE: \_\_\_\_\_

STAFF NAME: \_\_\_\_\_

LABEL NO. &amp; DATE: \_\_\_\_\_

Send this form to APA Subscription Claims, 750 First Street, NE, Washington, DC 20002-4242

PLEASE DO NOT REMOVE. A PHOTOCOPY MAY BE USED.



# Athletic Classmates, Physical Self-Concept, and Free-Time Physical Activity: A Longitudinal Study of Frame of Reference Effects

Ulrich Trautwein

Max Planck Institute for Human Development

Erin Gerlach

University of Bern

Oliver Lüdtke

Max Planck Institute for Human Development

The systematic analysis of factors that promote or impede physical activity in children is an urgent task for educational researchers. The present study investigated the reciprocal relationship between physical self-concept, teacher-assigned grades in physical education classes, and free-time physical activity, and analyzed positive and negative consequences of being in a class with high class-average physical ability. Data from a large, representative sample of 1,095 preadolescents from 66 classrooms were examined within a longitudinal framework. Multilevel analyses showed that membership in a class with high class-average physical ability was associated with lower physical self-concept and free-time physical activity and highlighted the significant role of teacher-assigned grades in the development of physical self-concept and physical activity. Furthermore, as predicted, there were positive reciprocal effects between physical self-concept and physical activity levels.

**Keywords:** physical ability, physical activity, self-concept, frame of reference effects, teacher-assigned grades

Over the past decade, there has been increasing interest in the physical activity levels of preadolescents and adolescents. A growing percentage of children in Western industrialized countries are overweight, increasing their risk of developing several medical problems, including cardiovascular disease and Type II diabetes (Brettschneider & Naul, 2004; Strauss & Pollack, 2001). Moreover, many researchers (e.g., Biddle & Mutrie, 2001; Sonstroem, 1984) assume that physical activity is associated with positive psychological and behavioral outcomes in children, such as global well-being, social integration, personality maturation, and positive self-views. Not surprisingly, there is broad consensus among researchers, physicians, educators, and politicians that the physical activity levels of children and adolescents need to be increased (e.g., National Association for Sport and Physical Education, 2004). The systematic analysis of factors associated with physical activity in children is thus a critical task for educational and developmental research.

Prior research with adolescents and adults has found evidence for a positive association between physical self-concept and physical activity levels. The present study addresses the reciprocal relationship between physical activity and physical self-concept in preadolescents, focusing on two additional factors. First, we examined the role of class composition in terms of physical ability levels. In academic subjects, class composition has been shown to be associated with frame of reference effects. Because students compare their own ability with that of their classmates, students in high-ability classes typically report comparatively low self-concepts (Marsh, 1987). We examined whether such frame of reference effects are also found in the physical domain, asking whether classmates' physical ability level is associated with individual physical self-concept and physical activity levels. Second, we sought to explore whether teacher-assigned grades account for any frame of reference effects found.

## Physical Self-Concept, Physical Ability, and Physical Activity in Children

In recent years, there has been an increasing awareness of the predictive effects of a student's self-concept for his or her feelings, motivation, activities, and achievement in various academic and nonacademic domains (e.g., Eccles, Wigfield, & Schiefele, 1998; Harter, 1998; Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2005; Trautwein, Lüdtke, Köller, & Baumert, 2006). The self-concept is a person's evaluation of his or her own qualities and standing in general or in specific domains. Students' self-concepts are typically measured via self-report measures (e.g., Byrne, 1996). In the developmental and educational literature, conceptions that center on global self-evaluations (e.g., "All in all, I am inclined to feel

---

Ulrich Trautwein and Oliver Lüdtke, Center for Educational Research, Max Planck Institute for Human Development, Berlin, Germany; Erin Gerlach, Institute of Sport Science, University of Bern, Berne, Switzerland.

The data presented are drawn from a large-scale German project directed by Wolf-Dietrich Brettschneider (Department for Sport and Health, University of Paderborn, Paderborn, Germany). The study was supported by a grant from the Foundation of the Sparkasse Bank of Paderborn. We are grateful to Herbert W. Marsh and Alexander Robitzsch for feedback on a draft of this article.

Correspondence concerning this article should be addressed to Ulrich Trautwein, Max Planck Institute for Human Development, Center for Educational Research, Lentzeallee 94, 14195 Berlin, Germany. E-mail: Ulrich.Trautwein@mpib-berlin.mpg.de, or Erin Gerlach, erin.gerlach@ispw.unibe.ch

that I am a failure"; Rosenberg, 1965) have gradually been replaced by models that focus on domain-specific self-evaluations (e.g., Marsh & Shavelson, 1985; Shavelson, Hubner, & Stanton, 1976) instead of or in addition to the global component. Typical domain-specific self-concept items are "I am quite good at mathematics" (math self-concept), "I have a poor vocabulary" (verbal self-concept), and "I'm talented when it comes to sports" (physical self-concept). Domain-specific academic self-concepts reflect a person's self-evaluation regarding a specific academic domain or ability and have been shown to be more closely related to academic outcomes than more global self-beliefs (e.g., Marsh, Trautwein, Lüdtke, Köller, & Baumert, 2006).

The conception of self-concept is related to the conception of self-efficacy beliefs (Bandura, 1997) but is not identical (Bong & Skaalvik, 2003). Whereas self-efficacy measures typically require a context-specific, goal-referenced evaluation of competence (e.g., "I am able to get an A on my next math test"), self-concept measures tap a person's generalized beliefs about his or her competence using normative statements.

Physical self-concept is known to be an important part of the self-definition in childhood (e.g., Bohrnstedt & Felson, 1983; Harter, 1998), and research has documented statistically significant relations between physical self-concept and physical ability (e.g., Sonstroem, 1997; Wigfield et al., 1997). In a recent study with 376 students in Grades 7, 8, and 9, Marsh, Chanal, and Sarrazin (2006) collected gymnastics self-concept and performance measures before (Time 1) and after (Time 2) a 10-week high school gymnastics program. As predicted by the authors, when performance at Time 1 was controlled, gymnastics self-concept positively predicted later performance to a statistically significant degree (standardized regression coefficient of  $\beta = .15$ ); when Time 1 gymnastics self-concept was controlled, gymnastics performance positively predicted later gymnastics self-concept ( $\beta = .15$ ). In other words, gymnastics performance and gymnastics self-concept were reciprocally related. At a casual glance, these effects may seem to be of moderate size. However, beta coefficients for cross-lagged effects (i.e., effects of one variable measured at an earlier time point on another variable measured at a later time point) in the range of .10 to .15 are common in real-world, nonexperimental longitudinal research on personal characteristics and can be considered meaningful (e.g., Roberts, Caspi, & Moffitt, 2003). When interpreting such cross-lagged effects, it is important to bear in mind that changes in physical self-concept, physical activity, and physical abilities are, of course, multiply determined (Ahadi & Diener, 1989) and that cross-lagged effects are potentially cumulative over time: The specific effect of a small beta coefficient may be quite substantial if it continues over extended periods of time (Neyer & Asendorpf, 2001; Prentice & Miller, 1992).

Furthermore, physical self-concept is closely related to the time and effort that people of all ages invest in sport and exercise (for summaries, see Biddle & Mutrie, 2001; Marsh, 2002; Sonstroem, 1997). For instance, in a study of contextual motivation in physical education, Standage, Duda, and Ntoumanis (2003) found the perceived physical competence of the 328 participants (mean age = 13.56 years) to be closely associated with their intention to engage in leisure-time physical activity in the near future ( $r = .46$ ). According to Cohen's (1988) classification of correlation coefficients, this constitutes a moderate to large association. Physical

self-concept also emerged as an important predictor of physical activity in a sample of about 1,500 Grade 6 students examined by Eccles and Harold (1991). In this study, physical self-concept correlated with free time spent on sport ( $r = .47$  for girls and  $r = .44$  for boys). In fact, the correlation between physical self-concept and free-time involvement in sport was significantly higher than the correlation between math or English self-concept and free-time involvement in math or English. Marsh, Papaioannou, and Theodorakis (2006) studied reciprocal effects of physical self-concept and exercise behavior at the beginning (Time 1) and end (Time 2) of the school year in a large sample of 2,786 students from 200 physical education classes at different levels of schooling. When initial exercise behavior was controlled, physical self-concept statistically significantly predicted Time 2 exercise behavior (standardized regression weight of  $\beta = .17$ ), and exercise behavior statistically significantly predicted Time 2 self-concept ( $\beta = .10$ ).

### Frame of Reference Effects, Self-Concept, and Physical Activity

The association between frame of reference effects, physical self-concept, and physical activity is of specific interest in the present study. There is now abundant evidence to show that the development of students' domain-specific self-concepts in the academic domain is considerably associated with the achievement of others in their immediate environment (Marsh & Craven, 2002; Marsh, Trautwein, Lüdtke, Baumert, & Köller, 2007; Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006). Individuals tend to focus on people close to them when constructing their frames of reference for social comparisons (Skaalvik & Skaalvik, 2002; Tesser, 1988). Very often, classmates or colleagues provide the frame of reference.

In many developmental and educational contexts, frame of reference effects result in students developing comparatively low self-concepts in high-achieving groups and comparatively high self-concepts in low-achieving groups. Rather than forming an impression of their own abilities in comparison with a "typical student" of their own age, students focus on their immediate environment. Extending earlier work by Davis (1966; see also Schwarzer, Lange, & Jerusalem, 1982), Marsh (1984, 1987) coined the term "big-fish-little-pond effect" (BFLPE) to describe this finding. Figure 1a illustrates the assumptions of the BFLPE. Academic self-concept is assumed to be positively related to individual achievement (high-achieving students will have higher academic self-concepts than low-achieving students), but negatively related to school- or class-average achievement (the same students will have lower academic self-concepts in a class where the average achievement is high and higher academic self-concepts in a class where the average achievement is low). Extensive empirical support for this effect has been accumulated over the last two decades (e.g., Lüdtke, Köller, Marsh, & Trautwein, 2005; Marsh & Hau, 2003; Marsh, Köller, & Baumert, 2001). Unfortunately, in contrast to the statistical significance of the predictor variables, the effect sizes for class- or school-level effects have received little attention.

Does the BFLPE generalize to the physical domain? On the one hand, it seems plausible to assume that the mechanisms governing the development of academic self-concepts also apply to physical self-concept. In fact, relative to other subjects, there may be a



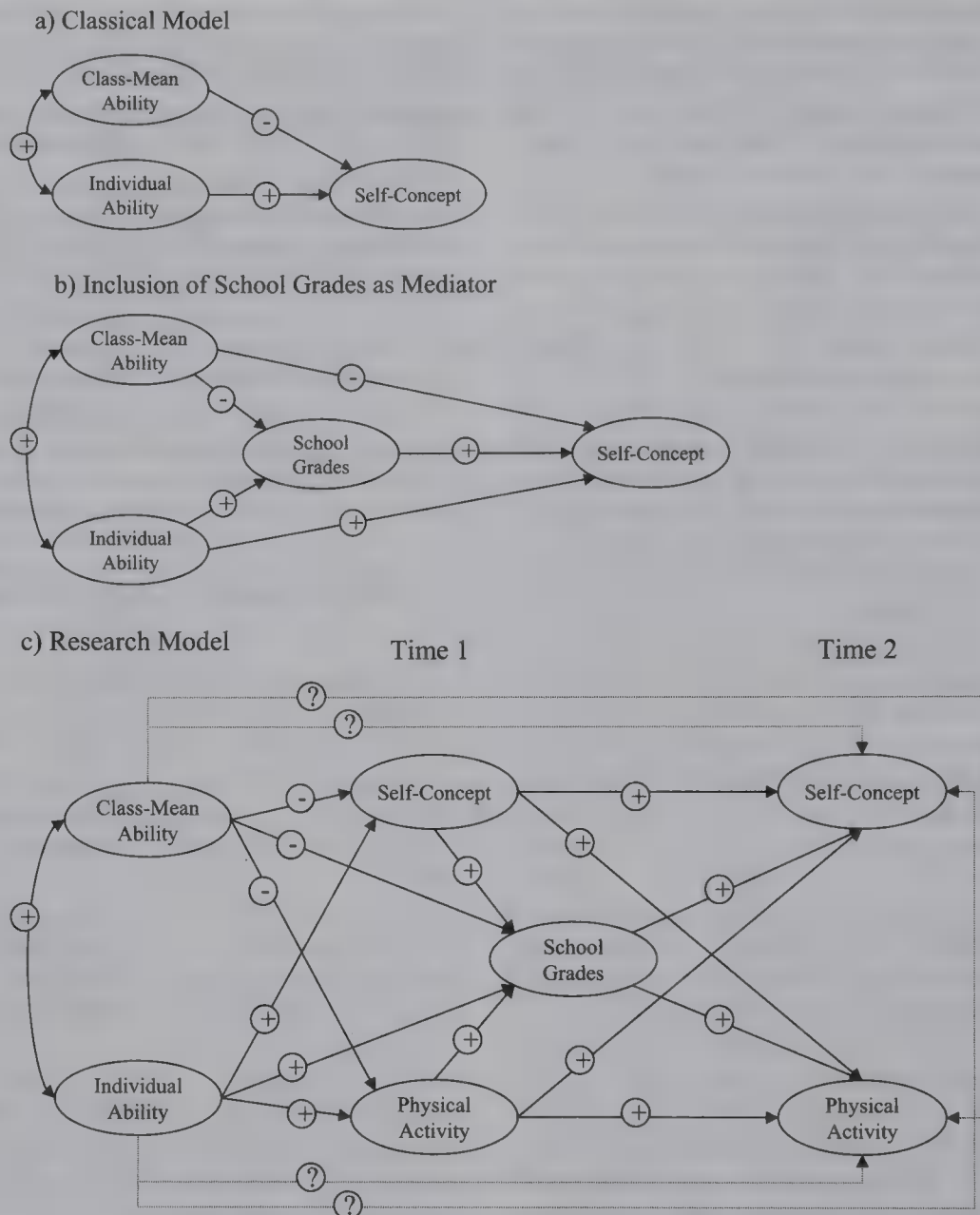


Figure 1. A schematic depiction of the big-fish-little-pond effect.

comparatively rich basis for social comparison in the physical domain. Students' abilities in the physical domain are more visible to their classmates than performance on a mathematics test, for example. Whereas students' oral contributions to lessons often provide the only direct basis for social comparison in mathematics (grades for written work typically are not directly accessible), performance in physical education (PE) lessons is always in the public sphere and may well elicit direct social comparisons (e.g., in athletics, team sports, or return matches). Moreover, it is common for ability rankings to be made explicit in PE lessons; for example, when teams are picked. It is still standard practice in many schools for teams to be selected by a process of peer nomination in PE lessons. Taken together, the specific features of PE lessons may reinforce the effects of social comparison processes.

On the other hand, physical self-concept is also more likely to be influenced by out-of-school factors than, say, mathematics or English self-concept. In addition, many sports involve improvement in an activity over time (e.g., running faster, jumping higher,

hitting a ball further); such improvements occur on an absolute metric and do not depend on external feedback. Clearly, out-of-school comparison as well as improvements in physical ability that can be evaluated on an absolute standard may counteract school-based social comparison processes.

It is interesting that, despite their theoretical and practical relevance, frame of reference effects on physical self-concept have remained virtually untested. As far as we are aware, the first study to investigate the BFLPE in the physical domain using the classical regression-based paradigm was reported by Chanal, Marsh, Sarrazin, and Bois (2005). These authors performed multilevel modeling with the responses of 405 Grade 7 and Grade 9 students from 20 groups participating in a 10-week gymnastics program. In line with the predictions of the BFLPE, individual achievement positively predicted gymnastics self-concept, whereas group-average gymnastics skills had a negative effect on gymnastics self-concept. Hence, the Chanal et al. (2005) study provides initial evidence that the BFLPE also holds in the context of physical education.

## Performance, Performance Feedback, and Frames of Reference

According to the theoretical assumptions of the BFLPE (e.g., Marsh & Craven, 2002), day-to-day social comparison with other students contributes to the emergence and development of the frame of reference effects that are encapsulated in the negative regression weight from class-average ability on self-concept. One of the greatest challenges to the BFLPE is the alternative hypothesis that the effect is not a consequence of social comparison processes, but simply mirrors teachers' grading practices (see Marsh, 1987; Trautwein & Lüdtke, 2005). This alternative explanation of the BFLPE rests on two assumptions: first, school grades—and not social comparison processes—are the main factor in self-concept development; second, average school grades are comparable in most schools, yielding similar average self-concepts despite diverging ability scores. Is there empirical support for this alternative explanation of the BFLPE?

Clearly, teacher-assigned grades are of pivotal importance for students' self-concepts (see Hansford & Hattie, 1982; Skaalvik & Skaalvik, 2002). Teacher-assigned grades provide students with transparent, relevant feedback and can directly impact their academic careers. Research has shown not only that grades impact students' self-concepts (e.g., Marsh et al., 2005; Valentine, DuBois, & Cooper, 2004), but also that academic self-concepts impact later grades (e.g., Trautwein, Lüdtke, Köller, & Baumert, 2006).

Grading practices may vary considerably within a school (Willingham, Pollack, & Lewis, 2002). One specific form of this variation is the "grading-on-a-curve" effect, which occurs when most teachers assign the best-performing student in the class a very good or good grade and the poorest-performing student a D or E grade. In this case, grades are assigned on a norm-referenced basis rather than a criterion-oriented basis (Marsh, 1987). Grading-on-a-curve effects are likely to vary across states and countries. In some countries, they are very large. For instance, in their study with more than 14,000 students from 621 schools in Germany, Trautwein, Lüdtke, Marsh, et al. (2006; Study 1) found an intra-class correlation coefficient (see Snijders & Bosker, 1999) of .63 for achievement in a standardized mathematics test. This means that 63% of the variance in mathematics achievement was explained by membership in different schools, indicating that the average achievement varied considerably across schools. In contrast, the intraclass correlation coefficient for teacher-assigned mathematics grades was only .05, indicating that the average teacher-assigned grade was remarkably similar across schools.

When teachers grade on a curve, the grades awarded only partially reflect students' objective academic standing. Given that students' academic self-concepts are markedly influenced by their grades, grading on a curve might introduce or reinforce the frame of reference effect postulated in the BFLPE model. Hence, grades are potentially a key mediator variable between individual and school-average achievement, on the one hand, and student motivation, on the other. This possibility is illustrated in Figure 1b. Whereas high individual achievement has a positive effect on school grades, high class-average achievement has a negative effect. In other words, students in classes with high average achievement receive lower grades for the same work. Grades, in turn, predict self-concept.

It is interesting that, despite their potential to reinforce the BFLPE, the role of school grades has rarely been systematically investigated. One exception is a study reported by Marsh (1987), who examined the association between school grades (GPA) and academic self-concepts in a sample of 2,213 grade 10 boys from 87 high schools. In this study, GPA significantly predicted academic self-concept and partially mediated the effect of individual achievement. The predictive effect of class-average achievement was only slightly affected by including GPA in the analyses. A particularly strong effect of school grades was reported by Trautwein, Lüdtke, Marsh, et al. (2006), who found mathematics school grades to have the single most important predictive effect on mathematics self-concept. However, the effects of individual and school-average mathematics achievement were only partly mediated by school grades.

Taken together, these two studies highlight the important role of school grades, but also point to frame of reference effects that cannot be fully explained by grading-on-a-curve effects. Better tests of the hypothesis that the BFLPE is strongly reinforced by grading practices are evidently needed; the present study attempts to provide such a test.

## The Present Investigation

At Time 1, a sample of 1,095 Grade 3 students from 66 classes participated in a test of physical fitness and completed questionnaires about their physical self-concept and free-time physical activity. At Time 2, approximately 15 months later, the same students again completed the questionnaire measures. The students were assigned their first grades for PE in the interval between Time 1 and Time 2.

Figure 1c illustrates the central assumptions of our research model. Our first set of hypotheses concerns frame of reference effects on physical self-concept and physical activity before and after the preadolescents in our sample were first given teacher feedback on their physical performance in the form of PE grades. We formulated three hypotheses in this regard. One, controlling for individual physical ability, we expected to find a negative predictive effect of class-average physical ability on students' physical self-concepts. Two, assuming that grading-on-a-curve effects were observable in our sample, we expected the negative predictive effect of class-average physical ability to be larger at Time 2, when PE grades had been awarded for the first time. Three, we tested for frame of reference effects with regards to physical activity. When their own physical ability is controlled, do students placed in a class with high class-average physical ability report more or less free-time physical activity? Both outcomes seem plausible. A student who considers most of her classmates to be better at PE might decide not to compete with them in that discipline, but to work hard on, say, mathematics. At the same time, the more athletic students there are in the class, the more likely it is that a student will be invited to join a sports club for social reasons. Against this background, our hypothesis of a negative relation between class-average physical ability and free-time physical activity was somewhat tentative.

Our second set of hypotheses concerned possible frame of reference effects on school grades. Given the evidence for grading-on-a-curve effects recently reported by Trautwein, Lüdtke, Marsh, et al. (2006; Study 1), we expected class-average physical ability



to negatively predict PE school grades: Students with the same physical ability were expected to receive lower grades when placed in a high-ability class than when placed in a low-ability class.

Our third set of hypotheses targeted the reciprocal relationship between physical self-concept and physical activity. We expected to find positive predictive effects of Time 1 physical self-concept on Time 2 physical activity and of Time 1 physical activity on Time 2 physical self-concept, when controlling for the respective Time 1 variables. At the same time, we hypothesized that these predictive effects would be largely mediated by PE grades.

Finally, we used gender and body fat (as indicated by the body mass index [BMI]) as additional predictor variables in all sets of analyses. These variables were mainly used as control variables. Based on the work by Eccles and Harold (1991; see also Marsh, 1993), we expected boys to report higher physical self-concepts and more free-time involvement in sports than girls. Furthermore, we expected body fat to have a negative impact on self-concept and physical activity; however, we were especially interested in whether BMI would have any effect on physical self-concept and activity when the actual performance level was controlled.

## Method

### Sample and Procedure

The data considered here derive from the Sport Involvement and Development in Children study conducted at the University of Paderborn, Germany (see Brettschneider & Gerlach, 2004, for more information). The present study draws on data collected from about one third of all Grade 3 students in the city of Paderborn (about 140,000 inhabitants), attending a total of 66 classes in 22 schools. Students in Germany are required to attend the elementary school located closest to the family's place of residence. To ensure that the sample was highly representative, schools were randomly selected for participation.

At Time 1 (T1; spring 2001), 1,185 Grade 3 students (50.7% female) were administered a physical ability test and a written questionnaire. The mean age at T1 was  $M = 9.67$  ( $SD = 0.64$ ). Approximately 15 months later (T2; summer 2002), the same children were re-contacted and asked to complete the questionnaire again. A total of 1,095 students (94.8% of the original sample) participated at T2. These 1,095 students represent the basis for the current study. The students who participated at both time points did not differ from those who dropped out in terms of physical ability, body fat, age, or gender (all  $ps > .10$ ). A statistically significant, albeit small, difference was found for physical self-concept, with dropouts ( $M = 3.27$ ,  $SD = 0.67$ ) reporting a higher physical self-concept than continuers,  $M = 3.11$ ,  $SD = 0.65$ ;  $t(1, 183) = 2.25$ ,  $p < .05$ , Cohen's  $d = 0.24$ .

The vast majority of participating students were Caucasian (>95%) and had German citizenship (83.8%). Children with Russian (6.7%) and Turkish citizenship (2.2%) were the largest minority groups. We computed the weight status of all participants using the German guidelines on obesity in childhood and adolescence (Kromeyer-Hausschild et al., 2001; for international guidelines see Cole, Bellizzi, Flegal, & Dietz, 2000). According to these guidelines, 76.2% of the students fell into the normal weight range, 12.4% were underweight, 7.4% were overweight, and 4.0% were obese.

In all classes, trained research assistants (mostly college students majoring in PE) administered the physical ability test during regular PE lessons. PE is a compulsory element of the elementary and high school curriculum in Germany; only students with severe chronic health issues that preclude participation are permanently excused from classes. All students in PE classes were required to participate in the physical ability test as part of a citywide program to survey the athletic abilities of elementary school children.

The questionnaire part of the study was administered in regular school hours within 9 weeks of the physical ability test. To guarantee high quality of the data collected, trained research assistants (university students) administered the questionnaire to small groups of about 6 students. Students were encouraged to ask for help if they did not understand a question.

### Instruments

*Standardized test of basic physical ability.* At T1, students' overall physical ability was tested by means of the Hagedorn Obstacle Course (Riepe & Zindel, 1999), a standardized test of physical coordination, balance, and speed. This parsimonious test measures complex coordination skills using apparatus to be found in any (German) sports hall. It has been externally validated by trainers, who provided expert ratings of various aspects of performance (e.g., flow of movement, appropriate use of strength, anticipation of the individual stations of the course). The course consists of a slalom, a balancing exercise, a forward roll, and getting over and through obstacles. Students completed the test twice; the average time needed to complete the course was taken as the student's physical ability score. To facilitate interpretation, physical ability scores were reverse scored such that higher scores indicate better physical skills. The validity of students' scores on this ability test is supported by significant correlations with observer ratings and other measures of physical fitness. In the present sample, for instance, the correlation with school PE grades was  $r = .44$  ( $p < .001$ ). In addition, physical ability test scores were associated with teachers' overall evaluation of the students' physical ability ( $r = .47$ ,  $p < .001$ ) and with BMI ( $r = -.41$ ;  $p < .001$ ).

*School grades.* At T2, students reported the PE grade they had been awarded at the end of Grade 3 using the six-level grading system implemented throughout Germany. We reverse coded these grades to produce the following six levels: *excellent* (6), *good* (5), *satisfactory* (4), *sufficient* (3), *poor* (2), and *very poor* (1).

*BMI.* The BMI is the measure most widely used to tap the percentage of body fat. The BMI is calculated by the formula

$$BMI = \frac{\text{weight}}{\text{height}^2},$$

where weight is measured in kilograms and height in meters. In the present study, height and weight were measured by trained research assistants.<sup>1</sup>

*Physical self-concept.* Five items from a German adaptation of the physical self-concept scale from Harter's (1985) Self Perception Profile for Children were administered ("I am very good at

<sup>1</sup> In preliminary analyses, the predictive power of the BMI was compared with a score that was standardized separately for boys and girls. The patterns of relations with the other variables under investigation proved to be almost identical in both approaches. We therefore used the original BMI approach.



sport"; "I learn quicker than others of my age in sport"; "I learn new exercises very quickly in sport"; "I am as good as others of my age in sport"; "I am just not good at sport") with a 4-point response format (ranging from 1 = *disagree* to 4 = *agree*). Internal consistency (Cronbach's  $\alpha$ ) was .76 at T1 and .83 at T2.

**Physical activity.** In Germany, all elementary and high school students have up to 3 hr of compulsory PE classes per week. These classes are taught by regular teachers, typically in the school's sports hall. Other than these regular PE classes, elementary schools offer limited opportunities for students to engage in physical activity. Rather, most physical and sports activities are organized by independent sports clubs (*Vereine*). These nonprofit organizations offer a variety of individual and team sports and are usually supported by the local authorities (which provide sports facilities and equipment) and the German Olympic Sport Federation. At both points of measurement, students reported how many hours they typically dedicated to physical activity at a sports club (*Verein*) for each day of the week separately. This response format has been shown to yield more reliable estimates than more global reports (e.g., Schwarz & Oyserman, 2001; Tourangeau, Rips, & Rasinski, 2000). Each student's responses were summed across the 7 days of the week to form an overall index of free-time physical activity.

### Statistical Analyses

We performed multilevel regression analyses to predict physical self-concept and physical activity at T1 and T2. In most studies conducted in schools, individual student characteristics are confounded with classroom or school characteristics because individual students are not randomly assigned to groups. This clustering effect introduces problems related to appropriate levels of analysis, aggregation bias, and heterogeneity of regression (Raudenbush & Bryk, 2002). For the present investigation, it is particularly important to note that the meaning of a variable at the student level may not bear any straightforward relation to its meaning at the classroom level. The negative BFLPE represents a dramatic example of this problem, in that achievement at the individual level is positively related to self-concept, whereas achievement at the class-average level may be unrelated or negatively related to self-concept. The juxtaposition of the effects of individual achievement and class-average achievement is inherently a multilevel issue that cannot be represented properly at either the individual or the classroom level. Particularly when major variables represent different levels, it is important to analyze data with appropriate multilevel statistical procedures. Multilevel modeling, a special form of regression analysis, provides a powerful methodology for handling hierarchical data and was used in this study. A detailed presentation of multilevel modeling (also referred to as hierarchical linear modeling; HLM) is beyond the scope of the present investigation and is available elsewhere (e.g., Raudenbush & Bryk, 2002; Snijders & Bosker, 1999).

In the present study, all multilevel analyses were computed using the computer program HLM 6 (Raudenbush, Bryk, Cheong, & Congdon, 2004). We specified three-level models, with students as the first level, classes as the second level, and schools as the third level. Predictor variables were included at Level 1 and Level 2 only; we decided to include class-average physical ability but not school-average physical ability in these analyses for three reasons

(for a similar rationale, see Trautwein, Lüdtke, Marsh, et al., 2006, Study 2). First, the correlation between class-average and school-average physical ability amounted to .79, yielding highly redundant information at the two levels; accordingly, multicollinearity would result if both variables were considered simultaneously. Second, from a theoretical point of view, class-average ability is more central to self-concept development and physical activity levels, given that it provides the most immediate frame of reference. Third, in hierarchical linear modeling, variables at lower levels can explain variance at higher levels, but not vice versa; hence, when class-average ability is included, this variable could also explain school-level differences that may exist.

To test our hypotheses (see above and Figure 1c), we specified several sets of multilevel models using HLM 6. As in ordinary regression analyses, one outcome variable was regressed on several predictor variables in each of these models. By specifying several consecutive models, researchers are able to observe the change in the predictive power of one variable when an additional variable is included. In this way, the associations between variables in a path model as well as mediator hypotheses such as that specified for PE grades can be tested within the framework of HLM. There are two main differences between our approach and typical studies using a structural equation modeling approach. First, multilevel regression analysis cannot test mediator hypotheses in a single step, but rather specifies a set of models; second, our analyses are based on manifest indicators and not on latent constructs.

HLM does not report standardized regression coefficients. To enhance the interpretability of the regression coefficients produced, we standardized ( $M = 0$ ,  $SD = 1$ ) all continuous variables before performing the multilevel analyses. Dichotomous variables were retained in their original metric. Physical ability was aggregated at the class level to form an index of the overall level of physical ability in the class (and was not restandardized). All models reported are random-intercept models estimated by the full maximum likelihood method. We assessed model fit using the deviance values provided by HLM, which can be regarded as a measure of lack of fit between model and data (Snijders & Bosker, 1999). Deviance values are not usually interpreted directly; rather, differences in deviance values are calculated for several models for the same data set. The difference in deviance between two models has a chi-square distribution with degrees of freedom equal to the difference in the number of parameters estimated. Because we used the full maximum likelihood method, the chi-square statistic can be used to evaluate the change in model fit when either a fixed or a random effect is added. Large values of the chi-square statistic indicate that the model with more estimated parameters provides a better fit to the data than the more parsimonious model.

Effect sizes have found increasing use in educational research. The statistical significance of a finding says little about its substantive meaning or real-world importance (see Kline, 2004; Nickerson, 2000). Effect sizes allow the meaningfulness of an empirical result to be presented clearly and the findings of empirical studies to be more readily appreciated; they also help politicians and policy makers in their decision making (McCartney & Rosenthal, 2000; for a recent overview of effect sizes, see Grissom & Kim, 2005).

How can the meaningfulness of results from multilevel modeling be determined? In our study, we used three indicators of effect



size. First, in analogy to the measure of explained variance in ordinary linear regression models, we report the overall proportion of variance explained by the predictor variables for each model. This measure is determined by calculating the proportion of the total variance that is reduced when the predictor variables are introduced into the specific model (see Snijders & Bosker, 1999).

Second, we report easily interpretable regression coefficients for Level 1 variables. Because we standardized all continuous Level 1 predictor and outcome variables before entering them in our multilevel models, the coefficients of the continuous Level 1 variables can be interpreted in almost the same way as the standardized regression coefficients resulting from ordinary regression analysis. With reference to Cohen's (1988) suggestions for correlations, we consider a regression coefficient of 0.10 to mark the lower bound for a meaningful effect (for a similar rationale, see Roberts et al., 2003). Because gender was not standardized, the regression coefficients for gender show the differences in girls' and boys' outcome variables in standard deviations, controlled for the other predictor variables.

Third, in line with the majority of previous multilevel research (e.g., Marsh & Hau, 2003), we did not restandardize our Level 2 predictor variable (class-average physical ability). Hence, the class-level regression weights show change in the dependent variable (physical self-concept/activity) corresponding to an increase of one unit in class-average physical ability, expressed in the metric of physical ability at the student level. Tymms (2004) proposed that the effect size for continuous Level 2 predictors in multilevel models, which is comparable with Cohen's  $d$ , be calculated using the formula:

$$\Delta = 2 \times B \times SD_{\text{predictor}} / \sigma_e$$

where  $B$  is the unstandardized regression coefficient in the multilevel model,  $SD_{\text{predictor}}$  is the standard deviation of the predictor variable at the class level, and  $\sigma_e$  is the residual standard deviation at the student level. The resulting effect size describes the difference in the dependent variable between two classes that differ by two standard deviations on the predictor variable. Given the important role of physical activity for lifetime health and the association between physical self-concept and overall self-esteem, even a small impact of the class composition should be of interest to researchers and practitioners. We thus suggest that an effect size of  $\Delta \geq 0.20$  can be considered of practical significance in the present research.

## Results

### *Physical Self-Concept and Physical Activity: Descriptive Results*

Overall, students reported rather positive physical self-concepts at T1 ( $M = 3.11$ ,  $SD = 0.65$ ) and T2 ( $M = 3.04$ ,  $SD = 0.67$ ). The moderate decrease between T1 and T2 was statistically significant,  $t(1094) = 3.66$ ,  $p < .001$ ,  $d = 0.11$ , and is in line with research showing that self-concepts decline over the preadolescent years (e.g., Wigfield et al., 1997). Average BMI at T1 was 17.37 ( $SD = 2.94$ ). The correlation between T1 and T2 physical self-concept was  $r = .58$  ( $p < .001$ ). At both measurement points, the majority of students (T1: 57.3%; T2: 59.1%) reported attending sport clubs at least once a week. The average free time that these students

reported dedicating to sport amounted to  $M = 2.34$  hr per week ( $SD = 1.65$ ) at T1 and  $M = 2.39$  hr per week ( $SD = 1.78$ ) at T2. Overall mean free-time physical activity did not differ between T1 and T2,  $t(1094) = -0.74$ ,  $ns$ . The correlation between T1 and T2 physical activity was  $r = .47$  ( $p < .001$ ).

### *Predicting T1 Physical Self-Concept and Physical Activity: Evidence for the BFLPE*

Our first research question addresses frame of reference effects on physical self-concept and physical activity. Do the data indicate a BFLPE on the physical self-concepts of preadolescents before they are awarded their first PE grades? To answer this question, we first analyzed whether there was a difference in the average ability level of students in different classes by specifying an empty model in HLM. An empty model includes only the outcome variable, but no predictor variables, and indicates whether the mean of the outcome variable differs meaningfully across classes or schools. This was indeed the case: More than 15% of the total variance was located at the class or school levels.

We next used T1 physical self-concept as the dependent variable. Table 1 reports the results of these analyses. In model M1, we specified the classical BFLPE model as illustrated in Figure 1a by including both individual physical ability scores and class-average physical ability scores as predictor variables. As documented in Table 1, individual physical ability positively predicted physical self-concept; the regression coefficient of  $B = 0.45$  indicates that an increase in the ability score of one standard deviation was associated with an increase in physical self-concept of almost half a standard deviation. The statistically significant regression coefficient for class-average physical ability was  $B = -0.39$ . We calculated the effect size for this coefficient using the formula described above. With a standard deviation in class-average ability of  $SD = 0.44$  and a residual standard deviation at the student level of  $\sigma_e = 0.92$ , we found a small effect:

$$\Delta = 2 \times (-0.39) \times 0.439 / 0.92 = -0.37.$$

Hence, students with the same ability levels had higher physical self-concepts if they were in a class with low average ability than if their class showed high average ability. The overall percentage of variance explained amounted to 0.16. The model fit of model M1 was statistically significantly better than the fit of an empty model with physical self-concept as the dependent variable,  $\chi^2 = 191$ ,  $df = 4$ ,  $p < .001$ .<sup>2</sup>

In M2, we tested for the effects of gender and BMI. Because BMI and physical ability were meaningfully associated, we removed physical ability to estimate the unique effect of gender and BMI. As expected, both variables negatively predicted physical self-concept, but the percentage of explained variance was lower than in M1. M3 is a combination of M1 and M2, including ability as well as gender and BMI. The regression coefficients for indi-

<sup>2</sup> We additionally specified random effects for the impact of class-average achievement at the school level in model M1 as well as in models M4, M7, M10, and M14 to check whether class-average physical ability had the same effect on the outcome variables in all 22 schools. For all these analyses, the variance components were statistically nonsignificant (all  $ps > .10$ ).

Table 1  
*Predicting T1 Physical Self-Concept and T1 Physical Activity: Results From Multilevel Modeling*

Predictors	T1 self-concept						T1 physical activity					
	M1		M2		M3		M4		M5		M6	
	<i>B</i>	<i>SE</i>	<i>B</i>	<i>SE</i>	<i>B</i>	<i>SE</i>	<i>B</i>	<i>SE</i>	<i>B</i>	<i>SE</i>	<i>B</i>	<i>SE</i>
Intercept	0.00	0.03	0.67***	0.07	0.38***	0.08	-0.01	0.04	0.37**	0.12	0.25	0.12
Physical ability test: class average	-0.39***	0.06			-0.36***	0.06	-0.30**	0.11			-0.30**	0.11
Physical ability test: individual score	0.45***	0.02			0.40***	0.04	0.17***	0.03			0.17***	0.03
Gender: female			-0.44***	0.04	-0.25***	0.05			-0.24**	0.07	-0.16*	0.07
BMI			-0.16***	0.03	-0.01	0.03			-0.01	0.03	0.05	0.03
<i>R</i> <sup>2</sup>	0.16		0.08		0.18		0.03		0.02		0.04	
Deviance	2,915		3,021		2,896		3,060		3,072		3,049	
<i>df</i>	6		6		8		6		6		8	

Note. *B* = unstandardized regression coefficient; *SE* = standard error of *B*; BMI = Body Mass Index.

\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

vidual and class-average achievement decreased slightly relative to M1 when gender and BMI were entered at the same time. Whereas the predictive effect of gender remained significant (with male students reporting higher physical self-concepts) despite a decrease in the beta weight, BMI did not prove to have an independent predictive effect on physical self-concept after controlling for gender and ability. A total of 18% of the variance was explained in model M3. A comparison of M3 and M1 indicated that model fit of M3 was statistically significantly higher than that of M1,  $\chi^2 = 19$ ,  $df = 2$ ,  $p < .001$ .

Taken together, we found support for the predictions of the BFLPE in this sample of preadolescent students. It is important to note that participants in the present sample had not yet received school grades on their performance in PE; hence, the results cannot be attributed to grading effects.

We next applied the same approach to check for frame of reference effects on physical activity. In models M4 to M6, T1 free-time physical activity was used as the dependent variable instead of physical self-concept. In M4 (see Table 1), physical ability test scores at the individual and class levels were used as predictors. This model examines whether the frame of reference effect is restricted to physical self-concept (see M1) or also applies to physical activity. Results showed that individual physical ability was indeed associated with higher physical activity, whereas the class-average ability level negatively predicted physical activity. Students with the same physical ability reported less free-time physical activity when placed in high-ability classes. With a standard deviation in class-average ability of  $SD = 0.44$  and a residual standard deviation at the student level of  $\sigma_e = 0.96$ , we found an effect size of

$$\Delta = 2 \times (-0.30) \times 0.439/0.96 = -0.27.$$

Relative to the empty model with a deviance of 3,088 ( $df = 4$ ), the model fit of M4 was clearly improved,  $\chi^2 = 28$ ,  $df = 2$ ,  $p < .001$ .

When the ability indicators were substituted by gender and BMI (see M5), only gender significantly predicted physical activity. In accordance with findings of Eccles and Harold (1991), female students reported less physical activity than male students did. In M6, the ability variables as well as gender and BMI were used as predictors. In this analysis, the gender effect was somewhat

smaller than in M5. The regression coefficients for individual and class-average physical ability were of the same magnitude as in M4. The model fit of M6 was statistically significantly higher than that of M4,  $\chi^2 = 11$ ,  $df = 2$ ,  $p < .01$ . Controlling for the other predictor variables in M6, the effect size for class-average physical ability was  $\Delta = -0.27$ . In sum, frame of reference effects were not restricted to the physical self-concept, but also applied to free-time levels of physical activity.

### Predicting PE Grades

We next performed a set of analyses to predict PE grades. As illustrated in Figure 1b, we predicted that students with the same ability levels would be assigned lower grades in classes with high class-average physical ability than in classes with low class-average physical ability; that is, we expected to find evidence for grading on a curve. The results of these analyses are shown in Table 2. In model M7, individual and class-average physical ability were used as predictor variables. Individual physical ability positively predicted PE grades, whereas class-average physical ability had a negative predictive effect. The regression coefficient of  $B = -0.40$  corresponds to an effect size of  $\Delta = -0.42$ . Relative to an empty model with PE grades as dependent variable, the model fit of M7 was clearly improved,  $\chi^2 = 299$ ,  $df = 2$ ,  $p < .001$ . In model M8, gender and BMI were additionally introduced. Controlling for physical ability at the individual and class level, there was a small gender effect in favor of girls. BMI did not significantly predict PE grades.

Finally, in M9, we included T1 physical self-concept and physical activity as additional predictor variables. M9 tests the predictive effects of the T1 predictor variables on PE grades (see left to middle part of the research model in Figure 1c). The analysis showed that—given the same physical ability scores at T1—students' physical self-concept and free-time physical activity positively predicted their PE grades in a statistically significant way. With regression weights of  $B = 0.16$  and  $B = 0.11$ , however, the predictive effect of these two variables was considerably smaller than the predictive effect of the physical ability test score. When the additional predictor variables were included, class-average ability still statistically significantly predicted teacher-assigned



Table 2  
Predicting Teacher-Assigned PE Grades: Results From Multilevel Modeling

Predictors	M7		M8		M9	
	B	SE	B	SE	B	SE
Intercept	0.00	0.04	-0.18	0.10	-0.27**	0.09
Physical ability test: class average	-0.40***	0.11	-0.40***	0.11	-0.30**	0.11
Physical ability test: individual score	0.53***	0.04	0.53***	0.05	0.44***	0.04
Gender: female			0.14*	0.06	0.18**	0.06
BMI			-0.03	0.03	-0.04	0.04
T1 physical self-concept					0.16***	0.03
T1 physical activity					0.11***	0.02
R <sup>2</sup>	0.23		0.24		0.28	
Deviance	2,785		2,777		2,714	
df	6		8		10	

Note. PE grade = physical education school grade; B = unstandardized regression coefficient; SE = standard error of B; BMI = Body Mass Index.  
\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

grades, and the effect size was  $\Delta = -0.32$ . The model fit of M9 was a considerable improvement on that of M8,  $\chi^2 = 63$ ,  $df = 2$ ,  $p < .001$ .

#### Predicting T2 Physical Self-Concept and T2 Physical Activity

The last two sets of analyses were performed with T2 physical self-concept and T2 free-time physical activity as the outcome variables (see Figure 1c). We expected frame of reference effects to be visible in T2 physical self-concept and T2 physical activity, but we expected these effects to be greatly reduced or eliminated when the corresponding T1 variables were included. Moreover, we expected to observe a reciprocal relationship between physical self-concept and physical ability. In addition, we speculated that PE grades would function as a mediator variable between physical self-concept and physical activity. To test these predictions, we

used the well-established procedure of specifying a sequence of multilevel models in which additional predictor variables are included one after the other. The logic of the sequence follows our research model illustrated in Figure 1c.

In M10 (see Table 3), the BFLPE was replicated with T2 physical self-concept as the outcome variable instead of T1 physical self-concept as in M1. Individual ability positively predicted T2 physical self-concept. The regression coefficient amounted to 0.56, which was slightly higher than the corresponding coefficient in M1. Similarly, the regression coefficient for class-average physical ability amounted to -0.50, compared to -0.39 in M1, even though the ability measure was administered about 15 months before the T2 self-concept measure. With a residual standard deviation at the student level of  $\sigma_e = 0.86$ , the effect size for class-average physical ability amounted to  $\Delta = -0.51$ , which can be considered a moderate effect size. Relative to the empty model,

Table 3  
Predicting T2 Physical Self-Concept and T2 Physical Activity: Results From Multilevel Modeling

Predictors	T2 physical self-concept								T2 physical activity							
	M10		M11		M12		M13		M14		M15		M16		M17	
	B	SE	B	SE	B	SE	B	SE	B	SE	B	SE	B	SE	B	SE
Intercept	0.00	0.03	0.00	0.02	-0.06	0.06	0.02	0.06	0.00	0.04	0.00	0.03	-0.06	0.09	-0.04	0.09
Physical ability test: class average	-0.50***	0.06	-0.31***	0.05	-0.31***	0.05	-0.21**	0.05	-0.25***	0.10	-0.07	0.07	-0.08	0.07	-0.06	0.08
Physical ability test: individual score	0.56***	0.00	0.35***	0.02	0.34***	0.03	0.22***	0.03	0.17*	0.04	0.07*	0.03	0.10*	0.04	0.06	0.04
T1 physical self-concept			0.44***	0.03	0.44***	0.03	0.40***	0.03			0.05*	0.03	0.06*	0.03	0.04	0.02
T1 physical activity			0.06*	0.02	0.07**	0.02	0.03	0.02			0.45***	0.06	0.45***	0.06	0.44***	0.06
Gender					-0.04	0.04	-0.01	0.04					0.04	0.06	0.02	0.06
BMI					0.03	0.04	-0.02	0.03					0.06	0.03	0.07*	0.03
PE grade							0.28***	0.03							0.09*	0.04
R <sup>2</sup>	0.25		0.42		0.43		0.48		0.02		0.23		0.24		0.24	
Deviance	2,793		2,501		2,498		2,389		3,074		2,818		2,814		2,806	
df	6		8		10		11		6		8		10		11	

Note. B = unstandardized regression coefficient; SE = standard error of B; BMI = Body Mass Index; PE grade = physical education school grade.  
\*  $p < .05$ . \*\*  $p < .01$ . \*\*\*  $p < .001$ .

the model fit of M10 was much higher,  $\chi^2 = 314$ ,  $df = 2$ ,  $p < .001$ . Overall, M10 indicates that the relevance of frame of reference effects with respect to physical self-concept increased between T1 and T2.

In M11, we additionally included T1 physical self-concept and T1 physical activity as predictor variables. As illustrated in Figure 1c, we expected the effects of individual and class-average physical ability on T2 physical self-concept to be mediated considerably by T1 physical self-concept and T1 physical activity to have an independent predictive effect on T2 physical self-concept. In line with our expectations, T1 physical self-concept positively predicted T2 physical self-concept. The coefficient of 0.44 indicates moderate stability of physical self-concept between T1 and T2. It is interesting that we found statistically significant effects for all three remaining predictor variables. Individual ability had a positive effect: When controlling for T1 physical self-concept, the better the students' ability test score, the higher their T2 physical self-concept. Moreover, even when T1 self-concept was statistically controlled, class-average ability significantly negatively predicted T2 physical self-concept; the effect size was  $\Delta = -0.36$ . In other words, class-average ability predicted T1 physical self-concept (see M1) and additionally predicted T2 physical self-concept when controlling for T1 physical self-concept. Finally, T1 physical activity positively predicted T2 physical self-concept as hypothesized; however, at  $B = 0.06$ , the regression coefficient was small.

The pattern of results was very similar when we additionally controlled for gender and BMI (see M12). Controlling for the other variables, gender and BMI did not statistically significantly predict T2 physical self-concept. Moreover, the addition of these two variables did not improve model fit relative to M11,  $\chi^2 = 3$ ,  $df = 2$ , *ns*.

In M13, we additionally included PE grades as a predictor variable. As illustrated in Figure 1c, inclusion of PE grades allows us to examine whether this variable functions as a mediator between T1 predictor variables and T2 physical self-concept. In line with our hypothesis, PE grade statistically significantly predicted T2 physical self-concept. Moreover, PE grade mediated some of the predictive power of the other variables. The direct predictive effect of T1 physical activity was no longer statistically significant when PE grades were included. In addition, the size of the predictive effects of individual and class-average ability decreased; however, controlling for the other predictor variables, both individual ( $B = 0.22$ ) and class-average ability ( $B = -0.21$ ,  $\Delta = -0.26$ ) predicted T2 physical self-concept. The inclusion of PE grades in M13 led to a statistically significant improvement in model fit compared to M12,  $\chi^2 = 109$ ,  $df = 1$ ,  $p < .001$ .

Taken together, frame of reference effects already observable at T1 (see M1 in Table 1) not only were stable, but their total size substantially increased between T1 and T2 (see M10). PE grades proved to predict T2 physical self-concept and mediated some of the frame of reference effects (see M13). Physical activity predicted T2 physical self-concept, but this effect was mediated by PE grades.

Our final set of analyses was performed with T2 physical activity as the outcome variable (see Table 3). In M14, T2 physical activity was regressed on individual and class-average ability. Both variables statistically significantly predicted T2 physical activity ( $B = 0.17$  for individual ability,  $\Delta = -0.25$  for class-

average ability). Hence, although physical ability was measured 15 months prior to T2 physical activity, membership of a class with high class-average physical ability was associated with less free-time physical activity at T2. At  $\Delta = -0.23$ , the effect was small. The model fit of M14 was higher,  $\chi^2 = 27$ ,  $df = 2$ ,  $p < .001$ , than that of an empty model with T2 physical activity as the outcome variable.

T1 physical activity and T1 physical self-concept were entered as additional predictor variables in M15. As expected, T1 physical activity had a high regression weight, indicating a considerable amount of stability in that behavior. In line with our hypothesis, T1 physical self-concept positively predicted T2 physical activity; however, the regression coefficient was rather small. Similarly, the effect of individual ability was statistically significant, but of a small size. No statistically significant effect was found for class-average physical ability when T1 physical activity was controlled.

In M16, gender and BMI were used as additional predictor variables; these two variables did not significantly predict T2 physical activity once T1 physical activity was controlled. Moreover, the regression coefficients of the other predictors were only slightly affected when these two predictors were included. The model fit of M16 was not a significant improvement on that of M15,  $\chi^2 = 4$ ,  $df = 2$ , *ns*.

Finally, in M17, we also included PE grade as a predictor variable. PE grade proved to statistically significantly predict T2 physical activity. In addition, the direct effects of individual physical ability and T1 physical self-concept were no longer statistically significant when the PE grade was included. Relative to M16, model fit was statistically significantly improved,  $\chi^2 = 8$ ,  $df = 1$ ,  $p < .01$ .

Taken together, the analyses documented frame of reference effects on T2 physical activity (see M14), but these effects were already present at T1 and did not increase between T1 and T2. Physical self-concept and PE grades statistically significantly predicted T2 physical activity; the regression coefficients were of small size, however.

## Discussion

Using a longitudinal framework and a large sample of preadolescents, the present study documented frame of reference effects on both physical self-concept and physical activity, attested to the role of performance feedback in the development of the physical self-concept and physical activity levels, and provided evidence for reciprocal effects between physical self-concept and physical activity levels. The study has major implications for both theory and practice.

### *Frame of Reference Effects and the Role of Performance Feedback*

Frame of reference effects as expressed in the BFLPE are an important and still much-debated phenomenon (e.g., Dai, 2004; Marsh & Hau, 2003; Plucker et al., 2004) with far-reaching effects on children's psychosocial development and behavior. The present study replicated the basic predictions of the BFLPE, but it also extended this research tradition. First, our study indicates that the BFLPE is not restricted to classical academic domains, such as mathematics or languages, but also applies to the domain of



physical ability. Moreover, our findings indicate that the BFLPE can be observed in students as young as about 9 to 10 years of age. Hence, frame of reference effects seem to emerge early in students' careers.

Second, the main focus in BFLPE research is typically on self-concept as the dependent variable (see Marsh, 1991, and Trautwein, Lüdtke, Marsh, et al., 2006, for exceptions). In our study, free-time physical activity level was examined in addition to physical self-concept. Clearly, physical activity level is a behavioral outcome with high educational relevance (see Biddle & Mutrie, 2001). Given that physical activity level is an out-of-school outcome that is likely to be affected by numerous out-of-school factors (e.g., parents, opportunity structures), it is all the more noteworthy that we found evidence for school-based frame of reference effects. Hence, our finding of a BFLPE on free-time engagement in sports significantly adds to the existing literature on frame of reference effects.

Third, our study is one of the first to systematically examine the impact of school grades on the BFLPE. It has been argued that the BFLPE might simply reflect norm-referenced grading practices in which classmates are used as the frame of reference, or at least that the BFLPE is strengthened by these practices (see Marsh, 1987; Trautwein, Lüdtke, Marsh, et al., 2006). The present study contributes to this debate in three ways. First, although students had not yet received any school grades for their performance at the first measurement point, the BFLPE was already apparent. Hence, the results demonstrate that the BFLPE is not simply the product of grading practices. Second, in line with our predictions, we found frame of reference effects in PE grades: Students with the same ability levels received lower grades when placed in a high-ability class. Third, the BFLPE was stronger for T2 physical self-concept than for T1 physical self-concept. Class-average ability negatively predicted T2 physical self-concept in a statistically significant way, even when T1 physical self-concept—including the effect of class-average ability on T1 physical self-concept—was controlled. In our view, it is likely that the introduction of PE grades—which were themselves subject to frame of reference effects—at the end of Grade 3 strengthened the natural frame of reference effects that were already in operation. The finding that the predictive effect of class-average physical ability on T2 physical self-concept was largely reduced when PE grades were entered in the regression model supports this hypothesis.

### *Reciprocal Effects Between Physical Self-Concept and Physical Activity*

Based on theoretical models and prior research (e.g., Eccles & Harold, 1991; Marsh, Chanal, & Sarrazin, 2006), we had expected to find evidence for a reciprocal relationship between physical self-concept and physical activity. Our analyses indeed revealed statistically significant reciprocal effects over time. However, with regression coefficients of  $B = 0.05$  and  $B = 0.06$ , the size of the predictive effects was somewhat smaller than anticipated and smaller than the correlation between the two variables ( $r = .24$  at T1) might suggest. What might be possible explanations for these rather small regression coefficients? First, our sample consists of preadolescents. It is quite possible that reciprocal effects between physical self-concept and physical activity are more pronounced in older students, who are able to regulate their behavior more au-

tonomously. Second, both physical self-concept and physical activity are likely to be multiply determined. For instance, whether or not students engage in organized free-time physical activities might depend to some extent on appropriate opportunities being available in the neighborhood and on their academic achievement. Likewise, although physical activity in clubs is likely to be associated with increased physical ability (see M9 in Table 2), frame of reference effects may mean that this relationship does not translate directly into higher physical self-concepts. Students in sports clubs may compare their own ability with their classmates, but also with other members of their clubs. The latter frame of reference may be more demanding.

### *Practical Implications*

Does social comparison decrease students' self-concept and constitute an obstacle to physical activity? As indicated by the negative effect of class-average achievement on self-concept and physical activity (the negative BFLPE), frame of reference effects are likely to be associated with negative developments in physical self-concept and physical activity in many students. This finding has important practical implications, to the extent that frame of reference effects may constitute an unassailable obstacle for programs aiming to promote physical activity. From a health perspective, the goal is to promote a sufficient level of physical activity among all students. Given the pervasive power of frame of reference effects, there will always be students with low physical self-concepts in mixed-ability PE classes, and the activity levels of these students are likely to be negatively affected by frame of reference effects.

How can educators and physical activity programs deal with frame of reference effects? One approach might be a change in student evaluation practices. In a study with more than 2,000 students from 112 classes, Lüdtke et al. (2005) showed that students' self-concepts are enhanced if their teachers apply an individualized frame of reference when evaluating student achievement. Teachers with an individualized teacher frame of reference emphasize improvement relative to prior achievement, effort, and learning. However, the findings also indicated that negative frame of reference effects were not suppressed. Future studies need to identify kinds of teacher feedback that may be able to enhance the physical self-concept of all students and reduce the negative effect of high-achieving reference groups.

### *Limitations and Future Research*

Certain limitations of our study should be kept in mind when interpreting the results. First and foremost, our study suffers from a limitation facing practically all real-world, nonexperimental research: the possibility of third-variable explanations. For instance, we predicted that high class-average ability would be associated with comparatively low physical self-concept when controlling for individual ability, and we found support for this hypothesis. According to our model, the statistically significant regression coefficient of class-average ability indicates that social comparison processes have taken place. However, it is equally possible that other unobserved variables caused or contributed to this association. Given the longitudinal study design and the control for potentially important third variables (gender, BMI), our study



might be relatively robust against third-variable explanations. Unfortunately, there is no perfect solution to this third-variable problem in nonexperimental, real-world studies in general or in the present study in particular. Although experimental studies seem to mitigate some of the problems of nonexperimental studies, our research design is not easily translatable to an experimental approach for ethical reasons.

Second, our physical activity variable only covered the organized setting of sports clubs. Other free-time sports activities were not considered. We believe this approach to be well-suited for our study for two reasons. One, preadolescents in Germany typically become involved in free-time sports activities in the context of such sports clubs. As a result of increasing urbanization, industrialization, and traffic density, people no longer get much exercise in daily life. It is now the institutional rather than the informal sector that offers opportunities to exercise. In fact, more than half of our participants were active members of sports clubs. Two, physical activity in sports clubs is typically characterized by higher frequency, regularity, intensity, and commitment than are other out-of-school physical activities. Hence, we believe that our physical activity variable captures the most important aspects of out-of-school physical activity. Nevertheless, future studies might also include measures of nonorganized physical activity.

Third, our study focuses on physical self-concept as a crucial motivational determinant of physical activity. As documented in the analyses, self-concept had a considerable impact on physical activity, both concurrently and prospectively. However, we acknowledge that other related constructs may also have predictive power. For instance, the expectancy-value framework developed by Eccles and colleagues (Eccles, 1983; Eccles & Harold, 1991; Wigfield & Eccles, 2002) complements the expectancy (self-concept) component by the value component, which comprises intrinsic value, attainment value, utility value, and cost. The studies by Eccles and colleagues (e.g., Eccles & Harold, 1991; Wigfield et al., 1997) indicate that this value component also contributes to explaining physical activity. Hence, our prediction of physical activity might have profited from including a value component.

Fourth, in a similar vein, we purposely focused on school-related factors as central determinants of physical self-concept and physical activity outside schools and did not address the important socializing role of parents or nonschool peers, the sports club setting, or the effects of various opportunity structures (e.g., geographical proximity to sport clubs, etc.). However, we were able to establish a remarkable association between school factors and children's out-of-school behavior. Given that our longitudinal study was the first to examine frame of reference effects on physical activity, we believe this decision to be justified, but we acknowledge that more comprehensive models would be preferable in future research.

Fifth, prior research (Marsh & Hau, 2003) has shown that the BFLPE is a worldwide phenomenon. We do not see much reason why the BFLPE should not also be observed in PE classes worldwide (see also Chanal et al., 2005). It is not clear, however, whether our findings regarding frame of reference effects on free-time physical activity generalize across different school systems. It would be interesting to conduct cross-cultural research on this issue in countries or communities with different organizational structures. We suspect that frame of reference effects are particu-

larly strong in school systems and communities where there is a close link between school-based physical education and free-time physical activity.

To conclude, the present research highlights the significance of physical self-concept for physical activity levels, documents that high-ability classmates have a negative effect on both physical self-concept and physical activity, and attests to the significant role of performance feedback in the development of the physical self-concept and physical activity levels. We believe that our findings have important theoretical and practical implications, and would like to see them prompt research on ways to counter the negative effects that high-ability classmates have on physical activity.

## References

- Ahadi, S., & Diener, E. (1989). Multiple determinants and effect size. *Journal of Personality and Social Psychology*, 56, 398–406.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.
- Biddle, S. J. H., & Mutrie, N. (2001). *Psychology of physical activity*. London: Routledge.
- Bohrnstedt, G. W., & Felson, R. B. (1983). Explaining the relations among children's actual and perceived performances and self-esteem: A comparison of several causal models. *Journal of Personality and Social Psychology*, 45, 43–56.
- Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review*, 15, 1–40.
- Bracken, B. A. (Ed.). (1996). *Handbook of self-concept*. New York: Wiley.
- Brettschneider, W.-D., & Gerlach, E. (2004). *Sportengagement und Entwicklung im Kindesalter* [Sport involvement and development in children]. Aachen, Germany: Meyer & Meyer.
- Brettschneider, W.-D., & Naul, R. (2004). *Study on young people's lifestyles and sedentariness and the role of sport in the context of education and as a means of restoring the balance* (Final Report for the European Commission, Directorate-General for Education and Culture). Paderborn, Germany: University of Paderborn. Retrieved May 29, 2007, from <http://europa.eu.int/comm/sport/documents/lotpaderborn.pdf>
- Byrne, B. M. (1996). *Measuring self-concept across the life span: Issues and instrumentation*. Washington, DC: American Psychological Association.
- Chanal, J. P., Marsh, H. W., Sarrazin, P. G., & Bois, J. E. (2005). Big-fish-little-pond effects on gymnastics self-concept: Social comparison processes in a physical setting. *Journal of Sport & Exercise Psychology*, 27, 53–70.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cole, T. J., Bellizzi, M. C., Flegal, K. M., & Dietz, W. H. (2000). Establishing a standard definition for child overweight and obesity worldwide: International survey. *British Medical Journal*, 320, 1240–1243.
- Dai, D. Y. (2004). How universal is the big-fish-little-pond effect? *American Psychologist*, 59, 267–268.
- Davis, J. A. (1966). The campus as a frog pond: An application of the theory of relative deprivation to career decisions of college men. *American Journal of Sociology*, 72, 17–31.
- Eccles, J. S. (1983). Expectancies, values, and academic choice: Origins and changes. In J. Spence (Ed.), *Achievement and achievement motivation* (pp. 87–134). San Francisco: W. H. Freeman.
- Eccles, J. S., & Harold, R. D. (1991). Gender differences in sport involvement: Applying the Eccles' expectancy-value model. *Journal of Applied Sport Psychology*, 3, 7–35.
- Eccles, J. S., Wigfield, A., & Schiefele, U. (1998). Motivation to succeed.



- In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3, Social, emotional and personality development* (5th ed., pp. 1017–1094). New York: Wiley.
- Grissom, R. J., & Kim, J. J. (2005). *Effect sizes for research: A broad practical approach*. Mahwah, NJ: Erlbaum.
- Hansford, B. C., & Hattie, J. A. (1982). The relationship between self and achievement/performance measures. *Review of Educational Research*, 52, 123–142.
- Harter, S. (1985). *Manual for the Self-Perception Profile for Children*. Denver, CO: University of Denver.
- Harter, S. (1998). The development of self-representations. In W. Damon (Series Ed.) & N. Eisenberg (Vol. Ed.), *Handbook of child psychology: Vol. 3. Social, emotional, and personality development* (5th ed., pp. 553–617). New York: Wiley.
- Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.
- Kromeyer-Hausschild, K., Wabitsch, M., Kunze, D., Geller, F., Geiß, H. C., Hesse, V., et al. (2001). Perzentile für den Body-Mass-Index für das Kindes- und Jugendalter unter Heranziehung verschiedener deutscher Stichproben [Percentiles of body mass index in children and adolescents evaluated from different regional German studies]. *Monatsschrift Kinderheilkunde*, 149, 807–818.
- Lüdtke, O., Köller, O., Marsh, H. W., & Trautwein, U. (2005). Teacher frame of reference and the big-fish-little-pond effect. *Contemporary Educational Psychology*, 30, 263–285.
- Marsh, H. W. (1984). Self-concept: The application of a frame of reference model to explain paradoxical results. *Australian Journal of Education*, 28, 165–181.
- Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, 79, 280–295.
- Marsh, H. W. (1991). The failure of high-ability high schools to deliver academic benefits: The importance of academic self-concept and educational aspirations. *American Educational Research Journal*, 28, 445–480.
- Marsh, H. W. (1993). The multidimensional structure of academic self-concept: Invariance over gender and age. *American Educational Research Journal*, 30, 841–860.
- Marsh, H. W. (2002). A multidimensional physical self-concept: A construct validity approach to theory, measurement and research. *Psychology: The Journal of the Hellenic Psychological Society*, 9, 459–493.
- Marsh, H. W., Chanal, J. P., & Sarrazin, P. G. (2006). Self-belief does make a difference: A reciprocal effects model of the causal ordering of physical self-concept and gymnastics performance. *Journal of Sports Sciences*, 24, 101–111.
- Marsh, H. W., & Craven, R. (2002). The pivotal role of frames of reference in academic self-concept formation: The “big fish–little pond” effect. In F. Pajares & T. Urdan (Eds.), *Adolescence and education* (Vol. 2, pp. 83–123). Greenwich, CT: Information Age.
- Marsh, H. W., & Hau, K. (2003). Big-fish-little-pond effect on academic self-concept: A cross-cultural (26-country) test of the negative effects of academically selective schools. *American Psychologist*, 58, 364–376.
- Marsh, H. W., Köller, O., & Baumert, J. (2001). Reunification of East and West German school systems: Longitudinal multilevel modeling study of the big-fish-little-pond effect on academic self-concept. *American Educational Research Journal*, 38, 321–350.
- Marsh, H. W., Papaioannou, A., & Theodorakis, Y. (2006). Causal ordering of physical self-concept and exercise behavior: Reciprocal effects model and the influence of physical education teachers. *Health Psychology*, 25, 316–328.
- Marsh, H. W., & Shavelson, R. J. (1985). Self-concept: Its multifaceted, hierarchical structure. *Educational Psychologist*, 20, 107–125.
- Marsh, H. W., Trautwein, U., Lüdtke, O., Baumert, J., & Köller, O. (2007). The big-fish–little-pond effect: Persistent negative effects of selective high schools on self-concept after graduation. *American Educational Research Journal*, 44, 631–669.
- Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, 76, 397–416.
- Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2006). Integration of multidimensional self-concept and core personality constructs: Construct validation and relations to well-being and achievement. *Journal of Personality*, 74, 403–456.
- McCartney, K., & Rosenthal, R. (2000). Effect size, practical importance, and social policy for children. *Child Development*, 71, 173–180.
- National Association for Sport and Physical Education. (2004). *Moving into the future: National standards for physical education* (2nd ed.). Boston: McGraw-Hill.
- Neyer, F. J., & Asendorpf, J. B. (2001). Personality–relationship transaction in young adulthood. *Journal of Personality and Social Psychology*, 81, 1190–1204.
- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241–301.
- Plucker, J. A., Robinson, N. M., Greenspon, T. S., Feldhusen, J. F., McCoach, D. B., & Subotnik, R. F. (2004). It's not how the pond makes you feel, but rather how high you can jump. *American Psychologist*, 59, 268–269.
- Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin*, 112, 160–164.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Thousand Oaks, CA: Sage.
- Raudenbush, S. W., Bryk, A. S., Cheong, Y., & Congdon, R. T. (2004). *HLM 6: Hierarchical linear modeling*. Chicago: Scientific Software International.
- Riepe, L., & Zindel, M. (1999). *Talentsuche und Talentförderung in NRW* [Talent selection and promotion]. Düsseldorf, Germany: Ministerium für Arbeit, Soziales und Stadtentwicklung, Kultur und Sport des Landes Nordrhein-Westfalen.
- Roberts, B. W., Caspi, A., & Moffitt, T. (2003). Work experiences and personality development in young adulthood. *Journal of Personality and Social Psychology*, 84, 582–593.
- Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.
- Schwarz, N., & Oyserman, D. (2001). Asking questions about behavior: Cognition, communication, and questionnaire construction. *American Journal of Evaluation*, 22, 127–160.
- Schwarzer, R., Lange, B., & Jerusalem, M. (1982). Selbstkonzeptentwicklung nach einem Bezugsgruppenwechsel [Self-concept development after a reference-group change]. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 14, 125–140.
- Shavelson, R. J., Hubner, J. J., & Stanton, G. C. (1976). Validation of construct interpretations. *Review of Educational Research*, 46, 407–441.
- Skaalvik, E. M., & Skaalvik, S. (2002). Internal and external frames of reference for academic self-concept. *Educational Psychologist*, 37, 233–244.
- Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage.
- Sonstroem, R. J. (1984). Exercise and self-esteem. *Exercise and Sport Sciences Reviews*, 12, 123–155.
- Sonstroem, R. J. (1997). The physical self-system: A mediator of exercise and self-esteem. In K. R. Fox (Ed.), *The physical self: From motivation to well-being* (pp. 3–26). Champaign, IL: Human Kinetics.
- Standage, M., Duda, J. L., & Ntoumanis, N. (2003). A model of contextual motivation in physical education: Using constructs from self-determination and achievement goal theories to predict physical activity intentions. *Journal of Educational Psychology*, 95, 97–110.
- Strauss, R. S., & Pollack, H. A. (2001). Epidemic increase in childhood

overweight, 1986–1998. *Journal of the American Medical Association*, 286, 2845–2848.

Tesser, A. (1988). Toward a self-evaluation maintenance model of social behavior. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 21, pp. 181–227). San Diego, CA: Academic Press.

Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, England: Cambridge University Press.

Trautwein, U., & Lüdtke, O. (2005). The big fish little pond effect: Future research questions and educational implications. *Zeitschrift für Pädagogische Psychologie/German Journal of Educational Psychology*, 19, 137–140.

Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2006). Self-esteem, academic self-concept, and achievement: How the learning environment moderates the dynamics of self-concept. *Journal of Personality and Social Psychology*, 90, 334–349.

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology*, 98, 788–806.

Tymms, P. (2004). Effect sizes in multilevel models. In I. Schagen & K. Elliot (Eds.), *But what does it mean? The use of effect sizes in educa-*

*tional research* (pp. 55–66). London: National Foundation for Educational Research.

Valentine, J. C., DuBois, D. L., & Cooper, H. (2004). The relations between self-beliefs and academic achievement: A systematic review. *Educational Psychologist*, 39, 111–133.

Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In A. Wigfield & J. S. Eccles, (Eds.), *Development of achievement motivation* (pp. 173–195). San Diego, CA: Academic Press.

Wigfield, A., Eccles, J. S., Yoon, K. S., Harold, R. D., Arbretton, A., Freedman-Doan, K., & Blumenfeld, P. C. (1997). Changes in children's competence beliefs and subjective task values across the elementary school years: A three-year study. *Journal of Educational Psychology*, 89, 451–469.

Willingham, W., Pollack, J. M., & Lewis, C. (2002). Grades and test scores: Accounting for observed differences. *Journal of Educational Measurement*, 39, 1–37.

Received November 28, 2006

Revision received October 4, 2007

Accepted October 7, 2007 ■

**UNITED STATES POSTAL SERVICE® (All Periodicals Publications Except Requester Publications)**

**Statement of Ownership, Management, and Circulation**

1. Publication Title: *Journal of Educational Psychology*

2. Publication Number: 0022-0675

3. Filing Date: October 2008

4. Issue Frequency: Quarterly

5. Number of Issues Published Annually: 4

6. Annual Subscription Price: Indiv \$161

7. Complete Mailing Address of Known Office of Publication (Not printer) (Street, city, county, state, and ZIP+4®): 750 First Street, N.E., Washington, D.C. 20002-4242

8. Complete Mailing Address of Headquarters or General Business Office of Publisher (Not printer): 750 First Street, N.E., Washington, D.C. 20002-4242

9. Full Names and Complete Mailing Addresses of Publisher, Editor, and Managing Editor (Do not leave blank):

Publisher (Name and complete mailing address): American Psychological Association, 750 First Street, N.E., Washington, D.C. 20002-4242

Editor (Name and complete mailing address): Karen R. Harris, EdD, Vanderbilt University, Box 328 Peabody College, Nashville, TN 37203-5701

Managing Editor (Name and complete mailing address): Susan J.A. Harris, American Psychological Association, 750 First Street, N.E., Washington, D.C. 20002-4242

10. Owner (Do not leave blank. If the publication is owned by a corporation, give the name and address of the corporation immediately followed by the names and addresses of all stockholders owning or holding 1 percent or more of the total amount of stock. If not owned by a corporation, give the names and addresses of the individual owners. If owned by a partnership or other unincorporated firm, give its name and address as well as those of each individual owner. If the publication is published by a nonprofit organization, give its name and address.)

Full Name: American Psychological Association

Complete Mailing Address: 750 First Street, N.E., Washington, D.C. 20002-4242

11. Known Bondholders, Mortgagees, and Other Security Holders Owning or Holding 1 Percent or More of Total Amount of Bonds, Mortgages, or Other Securities. If none, check box: ☒ None

12. Tax Status (For completion by nonprofit organizations authorized to mail at nonprofit rates) (Check one):

The purpose, function, and nonprofit status of this organization and the exempt status for federal income tax purposes:

☒ Has Not Changed During Preceding 12 Months

☐ Has Changed During Preceding 12 Months (Publisher must submit explanation of change with this statement)

PS Form 3526, September 2007 (Page 1 of 8 (Instructions Page 3)) PSN 7530-01-000-9031 PRIVACY NOTICE: See our privacy policy on www.usps.com

13. Publication Title: *Journal of Educational Psychology*

14. Issue Date for Circulation Data Below: August 2008

15. Extent and Nature of Circulation

		Average No. Copies Each Issue During Preceding 12 Months	No. Copies of Single Issue Published Nearest to Filing Date
a. Total Number of Copies (Net press run)		3000	3000
b. Paid Circulation (By Mail and Outside the Mail)	(1) Mailed Outside-County Paid Subscriptions Stated on PS Form 3541 (Include paid distribution above nominal rate, advertiser's proof copies, and exchange copies)	2057	2161
	(2) Mailed In-County Paid Subscriptions Stated on PS Form 3541 (Include paid distribution above nominal rate, advertiser's proof copies, and exchange copies)		
	(3) Paid Distribution Outside the Mails Including Sales Through Dealers and Carriers, Street Vendors, Counter Sales, and Other Paid Distribution Outside USPS®	507	554
	(4) Bill Distribution by Other Classes of Mail Through the USPS (e.g., First-Class Mail®)		
c. Total Bill Distribution (Sum of 15b (1), (2), (3), and (4))		2564	2715
d. Free or Nominal Rate Distribution (By Mail and Outside the Mail)	(1) Free or Nominal Rate Outside-County Copies Included on PS Form 3541		
	(2) Free or Nominal Rate In-County Copies Included on PS Form 3541		
	(3) Free or Nominal Rate Copies Mailed at Other Classes Through the USPS (e.g., First-Class Mail)		
	(4) Free or Nominal Rate Distribution Outside the Mail (Carriers or other means)	221	219
e. Total Free or Nominal Rate Distribution (Sum of 15d (1), (2), (3), and (4))		221	219
f. Total Distribution (Sum of 15c and 15e)		2785	2934
g. Copies Not Distributed (See Instructions to Publishers #4 (page #3))		215	66
h. Total (Sum of 15f and g)		3000	3000
i. Percent Paid (15c divided by 15f times 100)		93%	93%

16. Publication of Statement of Ownership

☒ If the publication is a general publication, publication of this statement is required. Will be printed in the \_\_\_\_\_ issue of this publication.

☐ Publication not required.

17. Signature and Title of Editor, Publisher, Business Manager, or Owner

Signature: *Barbara Spruill*

Title: Dir. Service Center Operations

Date: 10/6/08

I certify that all information furnished on this form is true and complete. I understand that anyone who furnishes false or misleading information on this form or who omits material or information requested on the form may be subject to criminal sanctions (including fines and imprisonment) and/or civil sanctions (including civil penalties).

PS Form 3526, September 2007 (Page 2 of 8)



## Acknowledgment of Ad Hoc Reviewers

The editor of the *Journal* wishes to thank persons who have reviewed manuscripts.

Richard Allington Michelle Athanasiou	John W. Fantuzzo Jack M. Fletcher Jocelyn R. Folk Donna Y. Ford Emily Fox Tierra M. Freeman	Andreas Krapp Li-jen Kuo	Julie Rosenthal
Linda Baker James Baumann George G. Bear Tasha Beretvas Kathy S. Binder Sandra Leanne Bosacki Patricia Bowers Michelle Buehl Eric S. Buhs Stephen Burgess Deborah L. Butler Brian Byrne	David Galbraith Mollie Galloway Russell Gersten Richard C. Gilman Sandra Graham Anne Graves Keith T. Greaney Kathy E. Green Anne Gregory Frédéric Guay	Nonie K. Lesaux Christine Li-Grining Barbara G. Licht Janeann M. Lineman Robert F. Lorch Julie Lynch Martin F. Lynch	Farideh Salili Tanya Santangelo David Scanlon Marlene Scardamalia Wolfgang Schnotz Kathy L. Schuh Dale H. Schunk Paul A. Schutz Marjorie Seaton Priti Shah Timothy Shanahan Sungok Serena Shim Georgios D. Sideridis Gale M. Sinatra Rayne A. Sperling Jennifer F. Sullivan Jessica J. Summers Susan Swearer
Gian Vittorio Caprara Marilyn J. Chambliss Li Chieh Sandra Christenson Maria Chuy Tim J. Cleary Cathy Collins Block Vince Connelly Anne E. Cook Pierre Cormier Rhonda G. Craven Jennifer G. Cromley	R. Trent Haines Zach Hambrick Nicole Harlaar James Hartley John A.C. Hattie	Margo Mastropieri Adele W. Miccio Kouider Mokhtari Marjorie Montague Lesley Morrow David Moshman	Sha Tao
Heather A. Davis Peter F. de Jong	William Jeynes David W. Johnson Martin H. Jones Jennifer Joy	Jason Naile Susan B. Neuman Richard S. Newman	Peggy N. Van Meter Frank R. Vellutino
Linnea C. Ehri Joan L. Erickson	Stuart A. Karabenick Carol Anne Kardash Douglas F. Kauffman Devin Kearns Allison J. Kelaher Young John R. Kirby	Gabriele Oettingen Karen F. Osterman	Barbara Wasik Noreen Webb Rose-Marie Weber Philip H. Winne Lynda Brown Wright
		Reinhard H. Pekrun Raymond P. Perry Therese D. Pigott	Man Kate Xu
		Katherine Rawson K. Ann Renninger Cara Richards John T. E. Richardson Lloyd Rieber Karen Riley Alysia D. Roehrig	Steven R. Yussen
			Duan Zhang Akane Zusho



Volume 100  
Numbers 1-4

February–November 2008

Published quarterly  
by the  
American Psychological  
Association

# Journal of Educational Psychology

Margaret R. Harris, *Editor*  
Vanderbilt University

Eric M. Anderman, *Associate Editor*  
University of Kentucky

Donna M. Kulikowich, *Associate Editor*  
Pennsylvania State University

Victoria Miller  
University of Denver

Frank Pajares, *Associate Editor*  
University of Missouri

Jeffrey J. Walczyk, *Associate Editor*  
Louisiana Technical University

ISSN 0022-0663

Copyright © 2008 by the American Psychological Association  
750 First Street, NE, Washington, DC 20002-4242



### **Editor**

Karen R. Harris  
*Vanderbilt University*

### **Associate Editors**

Eric M. Anderman  
*University of Kentucky*

Jonna M. Kulikowich  
*Pennsylvania State University*

Gloria Miller  
*University of Denver*

Frank Pajares  
*Emory University*

Jeffrey J. Walczyk  
*Louisiana Technical University*

### **Advisory Editors**

Patricia Alexander  
Ellen R. Altermatt  
Lynley H. Anderman  
Robert Atkinson  
Carole Beal  
Hefer Bembenutty  
David A. Bergin  
Benita A. Blachman  
Mimi Bong  
Jere Brophy  
Scott W. Brown  
Adriana G. Bus  
Robert Calfee  
Joanne F. Carlisle  
Martha Carr  
Jerrell C. Cassidy  
Clark Chinn  
Namok Choi  
Donald L. Compton  
Alice J. Corkill  
H. Michael Crowson  
Anne E. Cunningham  
Teresa K. DeBacker  
Amanda M. Durik  
Pamela Beard El-Dinary  
Dorothy L. Espelage  
Jill Fitzgerald  
Douglas Fuchs  
Lynn S. Fuchs  
David C. Geary  
Alexandra Gottardo  
Steve Graham  
Barbara A. Greene

Charles R. Greenwood  
John Guthrie  
Douglas J. Hacker  
Vernon C. Hall  
Jenefer Husman  
Michael L. Kamil  
Avi Kaplan  
Robert M. Klassen  
Beth Kurtz-Costes  
Dan Lapsley  
Steve Lehman  
Willy Lens  
Joel R. Levin  
Elizabeth A. Linnenbrink  
Mary Lundeborg  
Charles MacArthur  
Linda M. Mason  
Richard E. Mayer  
Catherine McBride-Chang  
Valentina McInerney  
Debra K. Meyer  
Michael Middleton  
Lisa M. Soederberg Miller  
Raymond B. Miller  
Jens Möller  
Tamera B. Murdock  
Karen P. Murphy  
Darcia Narvaez  
Markku Niemivirta  
Jane Oakhill  
Rollanda E. O'Connor  
Richard Olson  
Helen Patrick

Nancy Perry  
Gary Phye  
Jan L. Plass  
Robert Reid  
Robert Renaud  
Alison M. Ryan  
Hollis S. Scarborough  
Christopher Schatschneider  
Wolfgang Schneider  
Marlene Schommer-Aikens  
Gregory Schraw  
Einar M. Skaalvik  
Susan Sonnenschein  
Laura M. Stapleton  
Joseph Stevens  
H. Lee Swanson  
John Sweller  
Sonya Symons  
Keith Thiede  
Theresa A. Thorkildsen  
Tim Urdan  
Ellen Usher  
Giovanni Valiante  
Sharon Vaughn  
Regina Völmeyer  
Charles A. Weaver III  
Kathryn R. Wentzel  
Allan Wigfield  
Joanna P. Williams  
Christopher A. Wolters  
Moshe Zeidner  
Barry J. Zimmerman

### **APA Journal Staff**

Susan J. A. Harris  
*Senior Director, Journals Program*

Clark Munsell  
*Account Manager*

Skip Maier  
*Director, Journal Services*

Annie Hill  
*Editorial Supervisor*

#### *Manuscript Editors*

Cara Bevington  
Jeffery Hume-Pratuch

Sharon Ramos  
*Editorial Assistant*

Patricia Beck  
Mark Winter

Micah Owino  
*Editorial Production Coordinator*

Paige W. Jackson  
*Director, Editorial Services*

Amy R. O'Keefe  
*Lead Editor*

# Author Index to Volume 100

## Key to Pagination

<i>Issue No.</i>	<i>Month</i>	<i>Pages</i>
1	February	1–234
2	May	235–490
3	August	491–726
4	November	727–1002

## ARTICLES

- Aikens, Nikki L., and Barbarin, Oscar—Socioeconomic Differences in Reading Trajectories: The Contribution of Family, Neighborhood, and School Contexts. . . . . 235
- Alexander, Joyce M.—*see* Neitzel, Carin
- Anderson, Carolyn J.—*see* Shim, S. Serena
- Andersson, Ulf—Mathematical Competencies in Children With Different Types of Learning Difficulties. . . . . 48
- Andrews, Glenda—*see* Hood, Michelle
- Baker, Eva L.—*see* Boscardin, Christy Kim
- Bandura, Albert—*see* Caprara, Gian Vittorio
- Barbaranelli, Claudio—*see* Caprara, Gian Vittorio
- Barbarin, Oscar—*see* Aikens, Nikki L.
- Barron, Kenneth E.—*see* Harackiewicz, Judith M.
- Barth, Amy E.—*see* Stuebing, Karla K.
- Baumert, Jürgen—*see* Klusmann, Uta
- Baumert, Jürgen—*see* Krauss, Stefan
- Blum, Werner—*see* Krauss, Stefan
- Boiché, Julie C. S., Sarrazin, Philippe G., Grouzet, Frederick M. E., Pelletier, Luc G., and Chanal, Julien P.—Students' Motivational Profiles and Achievement Outcomes in Physical Education: A Self-Determination Perspective. . . . . 688
- Boldrin, Angela—*see* Mason, Lucia
- Boscardin, Christy Kim, Muthén, Bengt, Francis, David J., and Baker, Eva L.—Early Identification of Reading Difficulties Using Heterogeneous Developmental Trajectories. . . . . 192
- Bove, Giannetta Del—*see* Caprara, Gian Vittorio
- Bradshaw, Catherine P.—*see* Koth, Christine W.
- Brand-Gruwel, Saskia—*see* Könings, Karen D.
- Broeck, Anja Van den—*see* Vansteenkiste, Maarten
- Broers, Nick J.—*see* Könings, Karen D.
- Brownell, Kelly D.—*see* Wang, Shirley S.
- Brunner, Martin—*see* Krauss, Stefan
- Caprara, Gian Vittorio, Fida, Roberta, Vecchione, Michele, Bove, Giannetta Del, Vecchio, Giovanni Maria, Barbaranelli, Claudio, and Bandura, Albert—Longitudinal Analysis of the Role of Perceived Self-Efficacy for Self-Regulated Learning in Academic Continuance and Achievement. . . . . 525
- Carbonneau, Noémie, Vallerand, Robert J., Fernet, Claude, and Guay, Frédéric—The Role of Passion for Teaching in Intrapersonal and Interpersonal Outcomes. . . . . 977
- Cerdán, Raquel, and Vidal-Abarca, Eduardo—The Effects of Tasks on Integrating Information From Multiple Documents. . . . . 209
- Chanal, Julien P.—*see* Boiché, Julie C. S.
- Cheng, Jacqueline H. S.—*see* Marsh, Herbert W.
- Cirino, Paul T.—*see* Stuebing, Karla K.
- Compton, Donald L.—*see* Fuchs, Lynn S.
- Conlon, Elizabeth—*see* Hood, Michelle
- Conrad, Nicole J.—From Reading to Spelling and Spelling to Reading: Transfer Goes Both Ways. . . . . 869
- Craddock, Caitlin F.—*see* Fuchs, Lynn S.
- Craddock, Caitlin—*see* Fuchs, Lynn S.
- Cutler, Laura, and Graham, Steve—Primary Grade Writing Instruction: A National Survey. . . . . 907
- Davis, Heather A., DiStefano, Christine, and Schutz, Paul A.—Identifying Patterns of Appraising Tests in First-Year College Students: Implications for Anxiety and Emotion Regulation During Test Taking. . . . . 942
- DeLeeuw, Krista E., and Mayer, Richard E.—A Comparison of Three Measures of Cognitive Load: Evidence for Separable Measures of Intrinsic, Extraneous, and Germane Load. . . . . 223
- DiStefano, Christine—*see* Davis, Heather A.
- Dresel, Markus—*see* Ziegler, Albert
- Durik, Amanda M.—*see* Harackiewicz, Judith M.
- Durik, Amanda M.—*see* Hulleman, Chris S.
- Easter, Marilyn—*see* Schommer-Aikins, Marlene
- Ehri, Linnea C.—*see* Rosenthal, Julie
- Elliot, Andrew J., and Murayama, Kou—On the Measurement of Achievement Goals: Critique, Illustration, and Application. . . . . 613
- Fernet, Claude—*see* Carbonneau, Noémie
- Fida, Roberta—*see* Caprara, Gian Vittorio
- Fletcher, Jack M.—*see* Fuchs, Lynn S.
- Fletcher, Jack M.—*see* Stuebing, Karla K.
- Forbes-Jones, Emma—*see* Moller, Arlen C.
- Francis, David J.—*see* Boscardin, Christy Kim
- Francis, David J.—*see* Stuebing, Karla K.
- Fuchs, Douglas—*see* Fuchs, Lynn S. (three entries)
- Fuchs, Lynn S., Compton, Donald L., Fuchs, Douglas, Hollenbeck, Kurstin N., Craddock, Caitlin F., and Hamlett, Carol L.—Dynamic Assessment of Algebraic Learning in Predicting Third Graders' Development of Mathematical Problem Solving. . . . . 829
- Fuchs, Lynn S., Fuchs, Douglas, Craddock, Caitlin, Hollenbeck, Kurstin N., Hamlett, Carol L., and Schatschneider, Christopher—Effects of Small-Group Tutoring With and Without Validated Classroom Instruction on At-Risk Students' Math Problem Solving: Are Two Tiers of Prevention Better Than One?. . . . . 491
- Fuchs, Lynn S., Fuchs, Douglas, Stuebing, Karla, Fletcher, Jack M., Hamlett, Carol L., and Lambert, Warren—Problem Solving and Computational Skill: Are They Shared or Distinct Aspects of Mathematical Cognition?. . . . . 30



- Furrer, Carrie—*see* Skinner, Ellen
- Ganschow, Leonore—*see* Sparks, Richard L.
- Gava, Monica—*see* Mason, Lucia
- Georgiou, George K., Parrila, Rauno, and Papadopoulos, Timothy C.—Predictors of Word Decoding and Reading Fluency Across Languages Varying in Orthographic Consistency. . . . 566
- Gerlach, Erin—*see* Trautwein, Ulrich
- Graham, Steve—*see* Cutler, Laura
- Graham, Steve—*see* Rogers, Leslie Ann
- Grouzet, Frederick M. E.—*see* Boiché, Julie C. S.
- Guay, Frédéric—*see* Carbonneau, Noémie
- Guglielmi, R. Sergio—Native Language Proficiency, English Literacy, Academic Achievement, and Occupational Attainment in Limited-English-Proficient Students: A Latent Growth Modeling Perspective. . . . 322
- Guthrie, John T.—*see* Klauda, Susan Lutz
- Hamlett, Carol L.—*see* Fuchs, Lynn S. (three entries)
- Harackiewicz, Judith M., Durik, Amanda M., Barron, Kenneth E., Linnenbrink-Garcia, Lisa, and Tauer, John M.—The Role of Achievement Goals in the Development of Interest: Reciprocal Relations Between Achievement Goals, Interest, and Performance. . . . 105
- Harackiewicz, Judith M.—*see* Hulleman, Chris S.
- Hau, Kit Tai—*see* Leong, Che Kan
- Hightower, A. Dirk—*see* Moller, Arlen C.
- Hollenbeck, Kurstin N.—*see* Fuchs, Lynn S. (two entries)
- Hood, Michelle, Conlon, Elizabeth, and Andrews, Glenda—Preschool Home Literacy Practices and Children's Literacy Development: A Longitudinal Analysis. . . . 252
- Hughes, Jan N., Luo, Wen, Kwok, Oi-Man, and Loyd, Linda K.—Teacher-Student Support, Effortful Engagement, and Achievement: A 3-Year Longitudinal Study. . . . 1
- Hughes, Jan N.—*see* Wu, Wei
- Hulleman, Chris S., Durik, Amanda M., Schweigert, Shaun A., and Harackiewicz, Judith M.—Task Values, Achievement Goals, and Interest: An Integrative Analysis. . . . 398
- Humbach, Nancy—*see* Sparks, Richard L.
- Husman, Jenefer—*see* Shell, Duane F.
- Jang, Hyungshim—Supporting Students' Motivation, Engagement, and Learning During an Uninteresting Activity. . . . 798
- Javorsky, James—*see* Sparks, Richard L.
- Jerman, Olga—*see* Swanson, H. Lee
- Johnson, Cheryl I.—*see* Mayer, Richard E.
- Johnson, Kathy E.—*see* Neitzel, Carin
- Jordan, Alexander—*see* Krauss, Stefan
- Kalyuga, Slava—When Less Is More in Cognitive Diagnosis: A Rapid Online Method for Diagnosing Learner Task-Specific Expertise. . . . 603
- Kiefer, Sarah M., and Ryan, Allison M.—Striving for Social Dominance Over Peers: The Implications for Academic Adjustment During Early Adolescence. . . . 417
- Kieffer, Michael J.—Catching Up or Falling Behind? Initial English Proficiency, Concentrated Poverty, and the Reading Growth of Language Minority Learners in the United States. . . . 851
- Kindermann, Thomas—*see* Skinner, Ellen
- Klauda, Susan Lutz, and Guthrie, John T.—Relationships of Three Components of Reading Fluency to Reading Comprehension. . . . 310
- Klusmann, Uta, Kunter, Mareike, Trautwein, Ulrich, Lüdtke, Oliver, and Baumert, Jürgen—Teachers' Occupational Well-Being and Quality of Instruction: The Important Role of Self-Regulatory Patterns. . . . 702
- Köller, Olaf—*see* Marsh, Herbert W.
- Könings, Karen D., Brand-Gruwel, Saskia, van Merriënboer, Jeroen J. G., and Broers, Nick J.—Does a New Learning Environment Come Up to Students' Expectations? A Longitudinal Study. . . . 535
- Koth, Christine W., Bradshaw, Catherine P., and Leaf, Philip J.—A Multilevel Study of Predictors of Student Perceptions of School Climate: The Effect of Classroom-Level Factors. . . . 96
- Krauss, Stefan, Brunner, Martin, Kunter, Mareike, Baumert, Jürgen, Blum, Werner, Neubrand, Michael, and Jordan, Alexander—Pedagogical Content Knowledge and Content Knowledge of Secondary Mathematics Teachers. . . . 716
- Kunter, Mareike—*see* Klusmann, Uta
- Kunter, Mareike—*see* Krauss, Stefan
- Kunter, Mareike—*see* Tsai, Yi-Miau
- Kwok, Oi-Man—*see* Hughes, Jan N.
- Lambert, Warren—*see* Fuchs, Lynn S.
- Landerl, Karin, and Wimmer, Heinz—Development of Word Reading Fluency and Spelling in a Consistent Orthography: An 8-Year Follow-Up. . . . 150
- Lau, Shun, and Nie, Youyan—Interplay Between Personal Goals and Classroom Goal Structures in Predicting Student Outcomes: A Multilevel Analysis of Person-Context Interactions . . . 15
- Leaf, Philip J.—*see* Koth, Christine W.
- Lee, Jihyun—*see* Stankov, Lazar
- Lemery-Chalfant, Kathryn—*see* Valiente, Carlos
- Lens, Willy—*see* Vansteenkiste, Maarten
- Leong, Che Kan, Tse, Shek Kam, Loh, Ka Yee, and Hau, Kit Tai—Text Comprehension in Chinese Children: Relative Contribution of Verbal Working Memory, Pseudoword Reading, Rapid Automatized Naming, and Onset-Rime Phonological Segmentation. . . . 135
- Liem, Arief Darmanegara—*see* Van Etten, Shawn
- Linnenbrink-Garcia, Lisa—*see* Harackiewicz, Judith M.
- Loh, Ka Yee—*see* Leong, Che Kan
- Loyd, Linda K.—*see* Hughes, Jan N.
- Lüdtke, Oliver—*see* Klusmann, Uta
- Lüdtke, Oliver—*see* Marsh, Herbert W.
- Lüdtke, Oliver—*see* Trautwein, Ulrich
- Lüdtke, Oliver—*see* Tsai, Yi-Miau
- Luo, Wen—*see* Hughes, Jan N.
- Marchand, Gwen—*see* Skinner, Ellen
- Marsh, Herbert W., Martin, Andrew J., and Cheng, Jacqueline H. S.—A Multilevel Perspective on Gender in Classroom Motivation and Climate: Potential Benefits of Male Teachers for Boys? . . . 78
- Marsh, Herbert W., Trautwein, Ulrich, Lüdtke, Oliver, and Köller, Olaf—Social Comparison and Big-Fish-Little-Pond Effects on Self-Concept and Other Self-Belief Constructs: Role of Generalized and Specific Others. . . . 510
- Martin, Andrew J.—*see* Marsh, Herbert W.
- Mason, Lucia, Gava, Monica, and Boldrin, Angela—On Warm Conceptual Change: The Interplay of Text, Epistemological Beliefs, and Topic Interest. . . . 291
- Mayer, Richard E., and Johnson, Cheryl I.—Revising the Redundancy Principle in Multimedia Learning. . . . 380
- Mayer, Richard E.—*see* DeLeeuw, Krista E.
- McInerney, Dennis M.—*see* Van Etten, Shawn

- Miller, Raymond B.—*see* Tabachnick, Sharon E.
- Moller, Arlen C., Forbes-Jones, Emma, and Hightower, A. Dirk—Classroom Age Composition and Developmental Change in 70 Urban Preschool Classrooms. . . . . 741
- Murayama, Kou—*see* Elliot, Andrew J.
- Muthén, Bengt—*see* Boscardin, Christy Kim
- Neitzel, Carin, Alexander, Joyce M., and Johnson, Kathy E.—Children's Early Interest-Based Activities in the Home and Subsequent Information Contributions and Pursuits in Kindergarten. . . . . 782
- Neubrand, Michael—*see* Krauss, Stefan
- Nie, Youyan—*see* Lau, Shun
- Nussbaum, E. Michael—Using Argumentation Vee Diagrams (AVDs) for Promoting Argument-Counterargument Integration in Reflective Writing. . . . . 549
- Papadopoulos, Timothy C.—*see* Georgiou, George K.
- Parrila, Rauno—*see* Georgiou, George K.
- Patton, Jon—*see* Sparks, Richard L.
- Pedersen, Sara—*see* Véronneau, Marie-Hélène
- Pelletier, Luc G.—*see* Boiché, Julie C. S.
- Pressley, Michael—*see* Van Etten, Shawn
- Reiser, Mark—*see* Valiente, Carlos
- Relyea, George E.—*see* Tabachnick, Sharon E.
- Rogers, Leslie Ann, and Graham, Steve—A Meta-Analysis of Single Subject Design Writing Intervention Research. . . . . 879
- Rosenthal, Julie, and Ehri, Linnea C.—The Mnemonic Value of Orthography for Vocabulary Learning. . . . . 175
- Ryan, Allison M., and Shim, S. Serena—An Exploration of Young Adolescents' Social Achievement Goals and Social Adjustment in Middle School. . . . . 672
- Ryan, Allison M.—*see* Kiefer, Sarah M.
- Ryan, Allison M.—*see* Shim, S. Serena
- Ryan, Richard M.—*see* Tsai, Yi-Miau
- Sanders, Elizabeth A.—*see* Vadasy, Patricia F.
- Sarrazin, Philippe G.—*see* Boiché, Julie C. S.
- Schatschneider, Christopher—*see* Fuchs, Lynn S.
- Schommer-Aikins, Marlene, and Easter, Marilyn—Epistemological Beliefs' Contributions to Study Strategies of Asian Americans and European Americans. . . . . 920
- Schutz, Paul A.—*see* Davis, Heather A.
- Schweigert, Shaun A.—*see* Hulleman, Chris S.
- Shell, Duane F., and Husman, Jenefer—Control, Motivation, Affect, and Strategic Self-Regulation in the College Classroom: A Multidimensional Phenomenon. . . . . 443
- Shim, S. Serena, Ryan, Allison M., and Anderson, Carolyn J.—Achievement Goals and Achievement During Early Adolescence: Examining Time-Varying Predictor and Outcome Variables in Growth-Curve Analysis. . . . . 655
- Shim, S. Serena—*see* Ryan, Allison M.
- Skinner, Ellen, Furrer, Carrie, Marchand, Gwen, and Kindermann, Thomas—Engagement and Disaffection in the Classroom: Part of a Larger Motivational Dynamic?. . . . . 765
- Soenens, Bart—*see* Vansteenkiste, Maarten
- Sparks, Richard L., Patton, Jon, Ganschow, Leonore, Humbach, Nancy, and Javorsky, James—Early First-Language Reading and Spelling Skills Predict Later Second-Language Reading and Spelling Skills. . . . . 162
- Stankov, Lazar, and Lee, Jihyun—Confidence and Cognitive Test Performance. . . . . 961
- Stanovich, Keith E.—*see* West, Richard F.
- Stoeger, Heidrun—*see* Ziegler, Albert
- Stuebing, Karla K., Barth, Amy E., Cirino, Paul T., Francis, David J., and Fletcher, Jack M.—A Response to Recent Reanalyses of the National Reading Panel Report: Effects of Systematic Phonics Instruction Are Practically Significant. . . . . 123
- Stuebing, Karla—*see* Fuchs, Lynn S.
- Swanson, H. Lee, Jerman, Olga, and Zheng, Xinhua—Growth in Working Memory and Mathematical Problem Solving in Children at Risk and Not at Risk for Serious Math Difficulties. . . . . 343
- Swanson, H. Lee—Working Memory and Intelligence in Children: What Develops?. . . . . 581
- Swanson, Jodi—*see* Valiente, Carlos
- Tabachnick, Sharon E., Miller, Raymond B., and Relyea, George E.—The Relationships Among Students' Future-Oriented Goals and Subgoals, Perceived Task Instrumentality, and Task-Oriented Self-Regulation Strategies in an Academic Environment. . . . . 629
- Tauer, John M.—*see* Harackiewicz, Judith M.
- Thijs, Jochem, and Verkuyten, Maykel—Peer Victimization and Academic Achievement in a Multiethnic Sample: The Role of Perceived Academic Self-Efficacy. . . . . 754
- Timmermans, Tinneke—*see* Vansteenkiste, Maarten
- Toplak, Maggie E.—*see* West, Richard F.
- Trautwein, Ulrich, Gerlach, Erin, and Lüdtke, Oliver—Athletic Classmates, Physical Self-Concept, and Free-Time Physical Activity: A Longitudinal Study of Frame of Reference Effects 988
- Trautwein, Ulrich—*see* Klusmann, Uta
- Trautwein, Ulrich—*see* Marsh, Herbert W.
- Trautwein, Ulrich—*see* Tsai, Yi-Miau
- Treat, Teresa A.—*see* Wang, Shirley S.
- Tremblay, Richard E.—*see* Véronneau, Marie-Hélène
- Tsai, Yi-Miau, Kunter, Mareike, Lüdtke, Oliver, Trautwein, Ulrich, and Ryan, Richard M.—What Makes Lessons Interesting? The Role of Situational and Individual Factors in Three School Subjects. . . . . 460
- Tse, Shek Kam—*see* Leong, Che Kan
- Vadasy, Patricia F., and Sanders, Elizabeth A.—Repeated Reading Intervention: Outcomes and Interactions With Readers' Skills and Classroom Instruction. . . . . 272
- Valiente, Carlos, Lemery-Chalfant, Kathryn, Swanson, Jodi, and Reiser, Mark—Prediction of Children's Academic Competence From Their Effortful Control, Relationships, and Classroom Participation. . . . . 67
- Vallerand, Robert J.—*see* Carbonneau, Noémie
- Vansteenkiste, Maarten, Timmermans, Tinneke, Lens, Willy, Soenens, Bart, and Broeck, Anja Van den—Does Extrinsic Goal Framing Enhance Extrinsic Goal-Oriented Individuals' Learning and Performance? An Experimental Test of the Match Perspective Versus Self-Determination Theory. . . . . 387
- Van Etten, Shawn, Pressley, Michael, McInerney, Dennis M., and Liem, Arief Darmanegara—College Seniors' Theory of Their Academic Motivation. . . . . 812
- van Merriënboer, Jeroen J. G.—*see* Könings, Karen D.
- Vecchio, Giovanni Maria—*see* Caprara, Gian Vittorio
- Vecchione, Michele—*see* Caprara, Gian Vittorio
- Verkuyten, Maykel—*see* Thijs, Jochem
- Véronneau, Marie-Hélène, Vitaro, Frank, Pedersen, Sara, and Tremblay, Richard E.—Do Peers Contribute to the Likelihood of Secondary School Graduation Among Disadvantaged Boys? . . . . . 429
- Vidal-Abarca, Eduardo—*see* Cerdán, Raquel
- Vitaro, Frank—*see* Véronneau, Marie-Hélène
- Wang, Shirley S., Treat, Teresa A., and Brownell, Kelly D.—Cognitive Processing About Classroom-Relevant Contexts: Teachers' Attention to and Utilization of Girls' Body Size, Ethnicity, Attractiveness, and Facial Affect. . . . . 473



West, Richard F., Toplak, Maggie E., and Stanovich, Keith E.—Heuristics and Biases as Measures of Critical Thinking: Associations with Cognitive Ability and Thinking Dispositions. . . . 930

West, Stephen G.—*see* Wu, Wei

Wimmer, Heinz—*see* Landerl, Karin

Wu, Wei, West, Stephen G., and Hughes, Jan N.—Effect of Retention in First Grade on Children’s Achievement Trajectories Over 4 Years: A Piecewise Growth Analysis Using Propensity Score Matching. . . . 727

Zheng, Xinhua—*see* Swanson, H. Lee

Ziegler, Albert, Dresel, Markus, and Stoeger, Heidrun—Addressees of Performance Goals. . . . 643

OTHER

Acknowledgment of Ad Hoc Reviewers . . . . . 1002

American Psychological Association Subscription Claims Information . . . . . 47, 416, 654, 987

Call for Nominations . . . . . 850

E-Mail Notification of Your Latest Issue Online! . . . . 149, 442, 602

Instructions to Authors . . . . . iii (February), 490, 726, ix (November)

Low Publication Prices for APA Members and Affiliates . . . . . 161, 386, 565, 941

Members of Underrepresented Groups: Reviewers for Journal Manuscripts Wanted . . . . . 66, 321, 509, 929

New Editors Appointed, 2010–2015 . . . . . 868

Subscription Order Form . . . . . 77, 309, ii (August), ii (November)

The main purpose of the *Journal of Educational Psychology* is to publish original, primary psychological research pertaining to education across all ages and educational levels. A secondary purpose of the *Journal* is the occasional publication of exceptionally important theoretical and review articles that are pertinent to educational psychology.

**Manuscript preparation.** Authors should prepare manuscripts according to the *Publication Manual of the American Psychological Association* (5th ed.). Manuscripts may be copyedited for bias-free language (see chap. 2 of the *Publication Manual*). Formatting instructions (all copy must be double-spaced) and instructions on the preparation of tables, figures, references, metrics, and abstracts appear in the *Manual*. For APA's Checklist for Manuscript Submission, see [www.apa.org/journals](http://www.apa.org/journals). **Abstract and keywords.** All manuscripts must include an abstract containing a maximum of 180 words typed on a separate page. After the abstract, please supply up to five keywords or brief phrases. **References.** References should be listed in alphabetical order. Each listed reference should be cited in text, and each text citation should be listed in the References. Basic formats are as follows:

- Haag, L., & Stern, E. (2003). In search of the benefits of learning Latin. *Journal of Educational Psychology*, 95, 174–178.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.
- Johnson, D. W., & Johnson, R. T. (1990). Cooperative learning and achievement. In S. Sharon (Ed.), *Cooperative learning: Theory and research* (pp. 173–202). New York: Praeger.

Adequate description of participants is critical to the science and practice of educational psychology; this allows readers to assess the results, determine generalizability of findings, and make comparisons in replications, extensions, literature reviews, or secondary data analyses. Authors should see guidelines for sample–subject description in the *Manual*. Appropriate indexes of effect size or strength of relationship should be incorporated in the results section of the manuscript (see pp. 5, 25–26 of the *Manual*). Information that allows the reader to assess not only the significance but also the magnitude of the observed effects or relationships clarifies the importance of the findings. **Figures.** Graphics files are welcome if supplied as Tiff, EPS, or PowerPoint. High-quality printouts or glossies are needed for *all* figures. The minimum line weight for line art is 0.5 point for optimal printing. When possible, please place symbol legends below the figure image instead of to the side. Original color figures can be printed in color at the editor's and publisher's discretion provided the author agrees to pay \$255 for one figure, \$425 for two figures, \$575 for three figures, \$675 for four figures, and \$55 for each additional figure.

**Publication policies.** APA policy prohibits an author from submitting the same manuscript for concurrent consideration by two or more publications. APA policy regarding posting articles on the Internet may be found at [www.apa.org/journals](http://www.apa.org/journals). In addition, it is a violation of APA Ethical Principles to publish "as original data, data that have been previously published" (Standard 8.13). As this is a primary journal that publishes original material only, APA policy prohibits publication of any manuscript or data that have already been published in whole or substantial part elsewhere. Authors have an obligation to consult journal editors concerning prior publication of any data on which their article depends. In addition, APA Ethical Principles specify that "after research results are published, psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use

such data only for that purpose, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release" (Standard 8.14). Authors must have available their data throughout the editorial review process and for at least 5 years after the date of publication.

**Masked review policy.** The *Journal* has a masked review policy, which means that the identities of both authors and reviewers are masked. Every effort should be made by the authors to see that the manuscript itself contains no clues to their identities. Authors should never use first person (*I, my, we, our*) when referring to a study conducted by the author(s) or when doing so reveals the authors' identities, e.g., "in our previous work, Johnson et al., 1998 reported that . . ." Instead, references to the authors' work should be in third person, e.g., "Johnson et al. (1998) reported that . . ." The authors' institutional affiliations should also be masked in the manuscript. Authors submitting manuscripts are required to include in the cover letter the title of the manuscript along with all authors' names and institutional affiliations. However, the first page of the manuscript should omit the authors' names and affiliations, but should include the title of the manuscript and the date it is submitted. Responsibility for masking the manuscript rests with the authors; manuscripts will be returned to the author if not appropriately masked. If the manuscript is accepted, authors will be asked to make changes in wording so that the paper is no longer masked. Authors are required to state in writing that they have complied with APA ethical standards in the treatment of their sample, or to describe the details of treatment. A copy of the APA Ethical Principles may be obtained at [www.apa.org/ethics/](http://www.apa.org/ethics/) or by writing the APA Ethics Office, 750 First Street, NE, Washington, DC 20002-4242. APA requires authors to reveal any possible conflict of interest in the conduct and reporting of research (e.g., financial interests in a test procedure, funding by pharmaceutical companies for drug research). Authors of accepted manuscripts will be required to transfer copyright to APA.

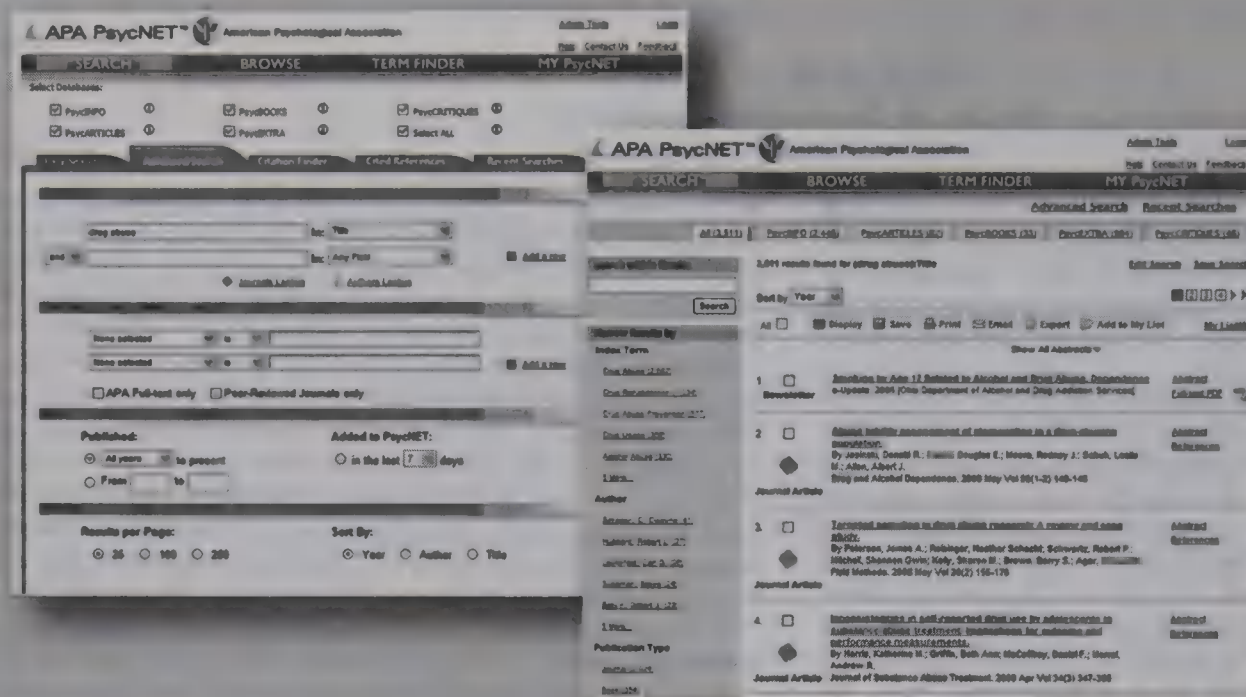
**Permissions.** Authors of accepted papers are required to obtain and provide to the editor on final acceptance all necessary permissions to reproduce in print and electronic form any copyrighted work, including, for example, test materials or portions thereof and photographs of people.

**Supplemental materials.** APA can now place supplementary materials online, which will be available via the published article in the PsycARTICLES database. To submit such materials, please see [www.apa.org/journals/authors/suppmaterial.html](http://www.apa.org/journals/authors/suppmaterial.html) for details. Authors of accepted papers will be asked to work with the editor and production staff to provide supplementary materials as appropriate.

**Submission.** Authors should submit their manuscripts electronically via the Manuscript Submission Portal at [www.apa.org/journals/edu](http://www.apa.org/journals/edu) (follow the link "Submit Manuscripts Electronically"). A checklist for manuscript submission, including guidelines for preparing the electronic file, can be found at [www.apa.org/journals/](http://www.apa.org/journals/). Correspondence regarding manuscripts should be sent to the Editor, Art Graessen, University of Memphis, Journal of Educational Psychology, 202 Psychology Building, Memphis, TN 38152-3230. In addition to addresses and phone numbers, authors should supply e-mail addresses, as most communications will be by e-mail. Fax numbers, if available, should also be provided for potential use by the editorial office and later by the production office. Authors should keep a copy of the manuscript to guard against loss. E-mail correspondence may be addressed to [jedgar@memphis.edu](mailto:jedgar@memphis.edu).

**Preparing files for production.** If your manuscript is accepted for publication, please follow the guidelines for file formats and naming provided at [www.apa.org/journals/authors/preparing\\_files.html](http://www.apa.org/journals/authors/preparing_files.html). If your manuscript was mask reviewed, please ensure that the final version for production includes a byline and full author note for typesetting.





## Maximize the Value of Your APA Membership

- Access over a century of psychological literature
- Discover millions of articles and book chapters in the social sciences
  - Browse the full text of APA journals from volume 1, issue 1
  - Download chapters from nearly 1,600 APA published books, including over 800 classic works

With **APA PsycNET®**, a world of behavioral sciences literature is at your fingertips. Choose from 3 database packages, each designed to meet specific professional needs:

- **APA PsycNET Silver** – access to the most current 3 years of journal literature
- **APA PsycNET Gold** – full access to all 5 APA research databases including more than 2.5 million abstracts in PsycINFO, full-text journal articles and book chapters, gray literature, reviews and more
- **APA PsycNET Platinum** – complete access to all APA electronic products and expanded licensing for one additional authorized user

Order now at **members.apa.org/access**  
and start enhancing your research or practice today!

*APA Databases...the psychology behind it all*

**Special  
APA  
member  
rates**





# Charles C Thomas

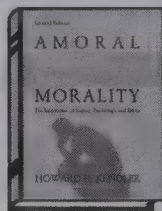
PUBLISHER • LTD.

P.O. Box 19265  
Springfield, IL 62794-9265

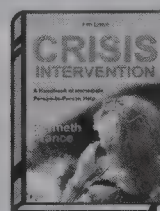
## Book Savings\*

Save 10% on 1 Book, Save 15% on 2 Books, Save 20% on 3 Books! (on separate titles only)

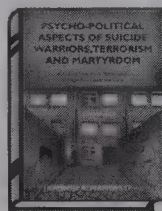
- Horovitz, Ellen G. & Sarah L. Eksten—**THE ART THERAPISTS' PRIMER: A Clinical Guide to Writing Assessments, Diagnosis, and Treatment.** '09, 272 pp. (7 x 10), 106 il., 2 Tables.
- Moon, Bruce L.—**EXISTENTIAL ART THERAPY: The Canvas Mirror.** (3rd Ed.) '09, 288 pp. (7 x 10), 51 il.
- Junge, Maxine Borowsky—**MOURNING, MEMORY AND LIFE ITSELF: Essays by an Art Therapist.** '08, 306 pp. (7 x 10), 38 il.
- Richard, Michael A., William G. Emmer, & William S. Hutchison, Jr.—**EMPLOYEE ASSISTANCE PROGRAMS: Wellness/Enhancement Programming.** (4th Ed.) '08, 348 pp. (8 x 10), 8 il., 1 table.
- Wiseman, Dennis G.—**THE AMERICAN FAMILY: Understanding Its Changing Dynamics and Place in Society.** '08, 152 pp. (7 x 10), 4 tables, paper.



- Kendler, Howard H.—**AMORAL THOUGHTS ABOUT MORALITY: The Intersection of Science, Psychology, and Ethics.** (2nd Ed.) '08, 270 pp. (7 x 10), \$59.95, hard, \$39.95, paper.



- France, Kenneth—**CRISIS INTERVENTION: A Handbook of Immediate Person-to-Person Help.** (5th Ed.) '07, 320 pp. (7 x 10), 3 il., \$65.95, hard, \$45.95, paper.



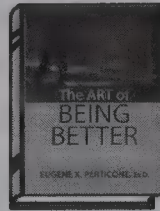
- Marvasti, Jamshid A.—**PSYCHO-POLITICAL ASPECTS OF SUICIDE WARRIORS, TERRORISM AND MARTYRDOM: A Critical View from "Both Sides" in Regard to Cause and Cure.** '08, 374 pp. (7 x 10), \$73.95, hard, \$53.95, paper.



- Horovitz, Ellen G.—**VISUALLY SPEAKING: Art Therapy and the Deaf.** '07, 250 pp. (7 x 10), 71 il., 5 tables, \$56.95, hard, \$36.95, paper.



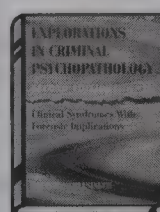
- Moon, Bruce L.—**INTRODUCTION TO ART THERAPY: Faith in the Product.** (2nd Ed.) '08, 226 pp. (7 x 10), 20 il., \$53.95, hard, \$33.95, paper.



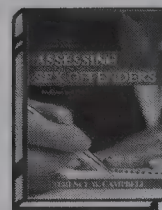
- Perticone, Eugene X.—**THE ART OF BEING BETTER: An Approach to Personal Growth.** '07, 268 pp. (7 x 10), \$58.95, hard, \$38.95, paper.



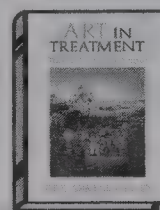
- Arrington, Doris Banowsky—**ART, ANGST, AND TRAUMA: Right Brain Interventions with Developmental Issues.** '07, 278 pp. (7 x 10), 123 il., (10 in color, paper edition only), \$63.95, hard, \$48.95, paper.



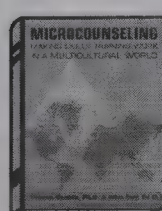
- Schlesinger, Louis B.—**EXPLORATIONS IN CRIMINAL PSYCHOPATHOLOGY: Clinical Syndromes With Forensic Implications.** (2nd Ed.) '07, 394 pp. (7 x 10), 3 il., 10 tables, \$79.95, hard, \$55.95, paper.



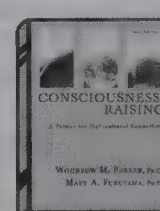
- Campbell, Terence W.—**ASSESSING SEX OFFENDERS: Problems and Pitfalls.** (2nd Ed.) '07, 376 pp. (7 x 10), 46 tables, \$74.95, hard, \$54.95, paper.



- Spring, Dee—**ART IN TREATMENT: Transatlantic Dialogue.** '07, 268 pp. (7 x 10), 39 il., 1 table, \$55.95, hard, \$37.95, paper.



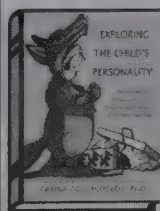
- Daniels, Thomas & Allen Ivey—**MICROCOUNSELING: Making Skills Training Work in a Multicultural World.** '07, 296 pp. (7 x 10), 12 il., 3 tables, \$65.95, hard, \$45.95, paper.



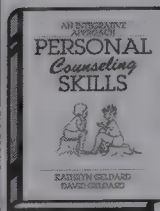
- Parker, Woodrow M. & Mary A. Fukuyama—**CONSCIOUSNESS-RAISING: A Primer for Multicultural Counseling** (3rd Ed.) '06, 264 pp. (7 x 10), 3 il., 5 tables, \$59.95, hard, \$39.95, paper.



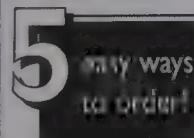
- Brooke, Stephanie L.—**THE CREATIVE THERAPIES AND EATING DISORDERS.** '08, 304 pp. (7 x 10), 20 il., 2 tables, \$64.95, hard, \$44.95, paper.



- Coulacoglou, Carina—**EXPLORING THE CHILD'S PERSONALITY: Developmental, Clinical and Cross-Cultural Applications of the Fairy Tale Test.** '08, 364 pp. (8 x 10), 22 il., 41 tables, \$78.95, hard, \$53.95, paper.



- Geldard, Kathryn & David Geldard—**PERSONAL COUNSELING SKILLS: An Integrative Approach.** '08, 316 pp. (7 x 10), 20 il., 3 tables, \$49.95, paper.



PHONE:  
1-800-258-8980  
or (217) 789-8980

FAX:  
(217) 789-9130

EMAIL:  
books@ccthomas.com  
Web: www.ccthomas.com

MAIL:  
Charles C Thomas  
Publisher, Ltd.  
P.O. Box 19265  
Springfield, IL 62794-9265

Complete catalog available at [ccthomas.com](http://ccthomas.com) • [books@ccthomas.com](mailto:books@ccthomas.com)

Books sent on approval • Shipping charges: \$7.75 min. U.S. / Outside U.S., actual shipping fees will be charged • Prices subject to change without notice

\*Savings include all titles shown here and on our web site. For a limited time only.

When ordering, please refer to promotional code **AMEPI108** to receive your discount.



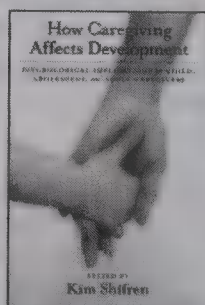
928005

# NEW RELEASES

from the American Psychological Association



AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION



## How Caregiving Affects Development

Psychological Implications for Child, Adolescent, and Adult Caregivers

*Edited by Kim Shifren*

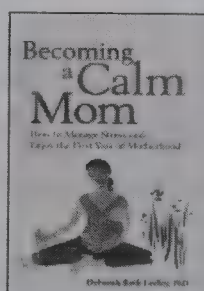
2009. 248 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0393-2

Item # 4318047



## An APA LifeTools book Becoming a Calm Mom

How to Manage Stress and Enjoy the First Year of Motherhood

*Deborah Roth Ledley, PhD*

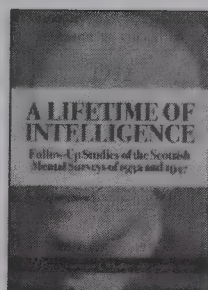
2009. 248 pages. Paperback.

List: \$19.95

APA Member/Affiliate: \$19.95

ISBN 978-1-4338-0404-5

Item # 4441012



## A Lifetime of Intelligence

Follow-Up Studies of the Scottish Mental Surveys of 1932 and 1947

*Ian J. Deary, Lawrence J. Whalley, and John M. Starr*

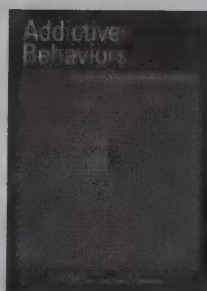
2009. 296 pages. Hardcover.

List: \$69.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0400-7

Item # 4318049



## Addictive Behaviors

New Readings on Etiology, Prevention, and Treatment

*Edited by G. Alan Marlatt and Katie Witkiewitz*

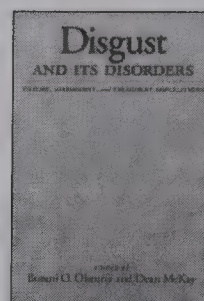
2009. 744 pages. Paperback.

List: \$49.95

APA Member/Affiliate: \$39.95

ISBN 978-1-4338-0402-1

Item # 4317166



## Disgust and Its Disorders

Theory, Assessment, and Treatment Implications

*Edited by Bunmi O. Olatunji and Dean McKay*

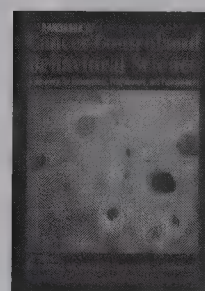
2009. 344 pages. Hardcover.

List: \$69.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0397-0

Item # 4316108



## Handbook of Cancer Control and Behavioral Science

A Resource for Researchers, Practitioners, and Policymakers

*Edited by Suzanne M. Miller, Deborah J. Bowen, Robert T. Croyle, and Julia H. Rowland*

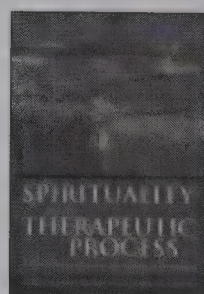
2009. 592 pages. Hardcover.

List: \$89.95

APA Member/Affiliate: \$59.95

ISBN 978-1-4338-0358-1

Item # 4317158



## Spirituality and the Therapeutic Process

A Comprehensive Resource From Intake to Termination

*Edited by Jamie D. Aten and Mark M. Leach*

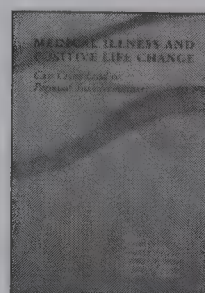
2009. 312 pages. Hardcover.

List: \$59.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0373-4

Item # 4317159



## Medical Illness and Positive Life Change

Can Crisis Lead to Personal Transformation?

*Edited by Crystal L. Park, Suzanne C. Lechner, Michael H. Antoni, and Annette L. Stanton*

2009. 280 pages. Hardcover.

List: \$69.95

APA Member/Affiliate: \$49.95

ISBN 978-1-4338-0396-3

Item # 4318048

AD0610

To Order: 800-374-2721 • [www.apa.org/books](http://www.apa.org/books)

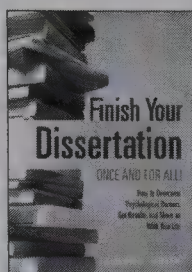


# NEW RELEASES

from the American Psychological Association



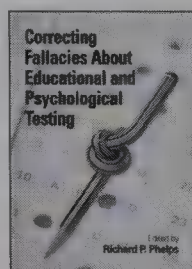
AMERICAN  
PSYCHOLOGICAL  
ASSOCIATION



## Finish Your Dissertation Once and for All!

How to Overcome Psychological Barriers,  
Get Results, and Move on With Your Life  
*Alison B. Miller*

2009. 264 pages. Paperback.  
List: \$29.95 • APA Member/Affiliate: \$24.95  
ISBN 978-1-4338-0415-1 • Item # 4313022



## Correcting Fallacies About Educational and Psychological Testing

*Edited by Richard P. Phelps*

2009. 296 pages. Hardcover.  
List: \$69.95 • APA Member/Affiliate: \$49.95  
ISBN 978-1-4338-0392-5 • Item # 4318046

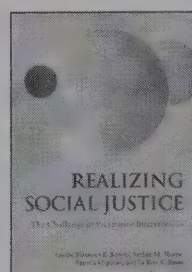


## The Duty to Protect

Ethical, Legal, and Professional  
Considerations for Mental Health  
Professionals

*Edited by James L. Werth, Jr.,  
Elizabeth Reynolds Welfel,  
and G. Andrew H. Benjamin*

2009. 272 pages. Hardcover.  
List: \$69.95 • APA Member/Affiliate: \$49.95  
ISBN 978-1-4338-0412-0 • Item # 4312013

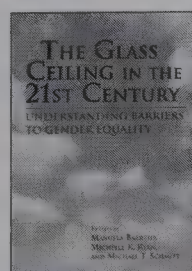


## Realizing Social Justice

The Challenge of Preventive Interventions

*Edited by Maureen E. Kenny,  
Arthur M. Horne, Pamela Orpinas,  
and LeRoy E. Reese*

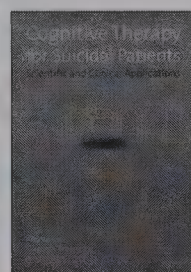
2009. 328 pages. Hardcover.  
List: \$69.95 • APA Member/Affiliate: \$49.95  
ISBN 978-1-4338-0411-3 • Item # 4317174



## The Glass Ceiling in the 21st Century Understanding Barriers to Gender Equality

*Edited by Manuela Barreto, Michelle K. Ryan,  
and Michael T. Schmitt*

2009. 344 pages. Hardcover.  
List: \$69.95 • APA Member/Affiliate: \$49.95  
ISBN 978-1-4338-0409-0 • Item # 4316109

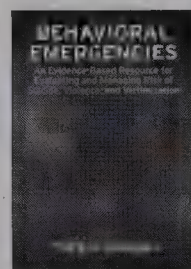


## Cognitive Therapy for Suicidal Patients

Scientific and Clinical Applications

*Amy Wenzel, Gregory K. Brown,  
and Aaron T. Beck*

2009. 386 pages. Hardcover.  
List: \$69.95 • APA Member/Affiliate: \$49.95  
ISBN 978-1-4338-0407-6 • Item # 4317169

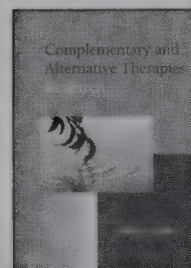


## Behavioral Emergencies

An Evidence-Based Resource  
for Evaluating and Managing Suicidal  
Behavior, Violence, and Victimization

*Edited by Phillip M. Kleespies*

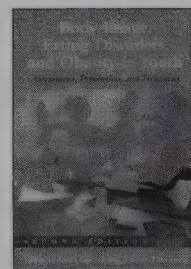
2009. 472 pages. Hardcover.  
List: \$69.95 • APA Member/Affiliate: \$49.95  
ISBN 978-1-4338-0406-9 • Item # 4317168



## Complementary and Alternative Therapies Research

*Tiffany Field*

2009. 200 pages. Hardcover.  
List: \$69.95 • APA Member/Affiliate: \$49.95  
ISBN 978-1-4338-0401-4 • Item # 4317165



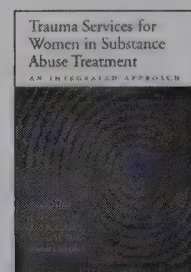
## Body Image, Eating Disorders, and Obesity in Youth

Assessment, Prevention,  
and Treatment

SECOND EDITION

*Edited by Linda Smolak  
and J. Kevin Thompson*

2009. 376 pages. Hardcover.  
List: \$59.95 • APA Member/Affiliate: \$49.95  
ISBN 978-1-4338-0405-2 • Item # 4317167

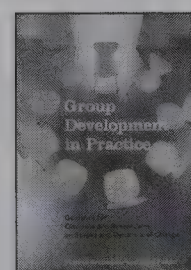


## Trauma Services for Women in Substance Abuse Treatment

An Integrated Approach

*Denise Hien, Lisa Caren Litt, Lisa R. Cohen,  
Gloria M. Miele, and Aimee Campbell*

2009. 304 pages. Hardcover.  
List: \$69.95 • APA Member/Affiliate: \$49.95  
ISBN 978-1-4338-0410-6 • Item # 4317173



## Group Development in Practice

Guidance for Clinicians and  
Researchers on Stages and  
Dynamics of Change

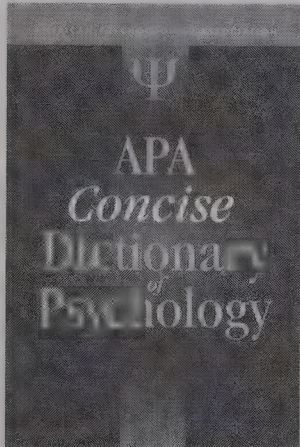
*Virginia Brabender and April Fallon*

2009. 312 pages. Hardcover.  
List: \$59.95 • APA Member/Affiliate: \$49.95  
ISBN 978-1-4338-0408-3 • Item # 4317172





## APA Concise Dictionary of Psychology



A handy abridgment of the award-winning guide to the language of the field from the world's largest professional association of psychologists, with many updates and revised definitions

- 10,000 entries offering clear and authoritative definitions—including many revised and updated definitions from the parent dictionary (*APA Dictionary of Psychology*)
- Balanced coverage across core areas of psychology—including cognitive, personality, social, developmental, health, clinical, experimental, neuropsychology, research methodology, and many others
- Thousands of incisive cross-references to deepen the user's understanding of related topics
- A "Quick Guide to Use" that explains stylistic and formal features at a glance
- Appendixes listing major figures in the history of psychology and psychological therapies and approaches
- Use as a standalone reference or as a portable alternative to the *APA Dictionary of Psychology*

2009 | 592 pages | Hardcover | List: \$39.95 | APA Member/Affiliate: \$39.95

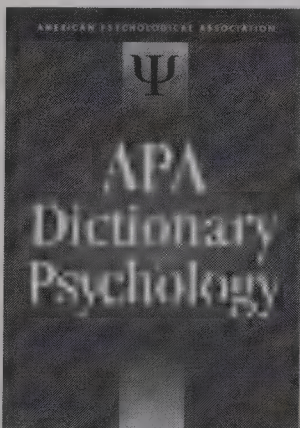
ISBN 978-1-4338-0391-8 | Item # 4311009

**RUSA Outstanding Reference Source**

**NYLA Best of Reference**

**Choice Outstanding Academic Title!**

## APA Dictionary of Psychology



With over 25,000 terms and definitions, the *Dictionary* encompasses all areas of research and application, and includes coverage of concepts, processes, and therapies across all the major subdisciplines of psychology. Ten years in the making and edited by a distinguished editorial board of nearly 100 psychological scholars, researchers and practitioners, the *APA Dictionary of Psychology* is destined to become the **most authoritative** reference of its kind.

*"This is an excellent reference work that should be ordered by every institutional library and every psychology department to make it accessible to students, faculty, researchers, and practitioners in all social science and health-related arenas."*

— **PsycCRITIQUES**

*"For the world's largest and most well known association of psychologists to produce a dictionary gives the work an 'out of the box' authority and prestige that few works share... Summing up: Essential."* — **Choice**

*"Ten years in the making, the American Psychological Association's (APA) new dictionary was well worth the wait."* — **Library Journal**, starred review

2006 | 1,024 pages | Hardcover | List: \$59.95 | APA Member/Affiliate: \$49.95

ISBN 978-1-59147-380-0 | Item # 4311009











